DELL Technologies

ベスト プラクティス

Dell EMC PowerFlex：ネットワーキングのベスト プラクティスと設計上の考慮事項

PowerFlexバージョン3.5.x

要約

本書では、Dell EMC PowerFlex™ソフトウェアデファインド ストレージで中核となる概念について説明しています。また、PowerFlexシステムに適したネットワークの設計、トラブルシューティングおよび保守のベストプラクティスを紹介しています。ベストプラクティスには単一サイトでの導入、レプリケーションを使用した複数サイトでの導入の両方について紹介しています。

2021年4月

H18390.3

改訂履歴

日付	説明
2021年4月	仮想ネットワークと動的ルーティングについて更新
2021年1月	インクルーシブな用語についての免責事項を追加
2020年6月	PowerFlex 3.5のリリースとブランド名変更 – レプリケーションについて修正および更新
2019年5月	VxFlex OS 3.0のリリース – 追記および更新
2018年7月	VxFlex OSのブランド名変更と全般的な修正 – VXLANに関する記載を追加
2016年6月	LAGに関する記載を追加
2015年11月	初版

謝辞

コンテンツ所有者：Brian Dean（ストレージ テクニカル マーケティング）

サポート：Neil Gerren、Igal Moshkovich、Matt Hobbs、Dan Aharoni、Rivka Matosevich

目次

概要

Dell EMC™ PowerFlex™ファミリーの製品には、PowerFlexソフトウェアデファインド ストレージが搭載されています。これはスケールアウト ブロック ストレージ サービスであり、予測可能なハイ パフォーマンスと大規模な耐障害性により、柔軟性、弾性、シンプル性を実現するよう設計されています。PowerFlexストレージ ソフトウェア（旧VxFlex OS）は、複数のOSおよびハイパーバイザー機能による各種の導入オプションに対応しています。

PowerFlexファミリーは、現在、ラックレベルの製品および2ノードレベルの製品で構成されています（アプライアンスとReady Node）。本書では、主にストレージ仮想化ソフトウェア レイヤー自体に焦点を当てているため、Ready Nodeに関する記載が大部分を占めますが、PowerFlexベースのストレージ システムに必要なネットワーク構築について理解を深めたい方にも適した内容になっています。

PowerFlexラックは、完全にモダン データセンター向けに設計されたラックスケール システムです。ラック ソリューションでは、ネットワークは事前に構成、最適化が行われているため、設計はPowerFlex Manager（PFxM）によって規定、実装、メンテナンスが行われます。本書では、ラック導入については記載していません。その他PowerFlexファミリーのソリューションについては、適切なネットワークの設計と実装が必要になります。PFxM 3.6リリースからは、アプライアンスの場合、特定の基準を満たし、PFxMで導入するトポロジーと一致するよう設定されていれば、サポート対象外の市販のスイッチを使用することもできます。これについては後述します。

PowerFlexの導入を正常に行えるかは、ネットワーク トポロジーが十分に設計されているかにかかっています。本書ではネットワークの選択について、またネットワークの選択がPowerFlexの各コンポーネントでどのようにトラフィック タイプと関連しているかについて説明しています。ソフトウェア バージョン3.5で導入されたPowerFlexネイティブ非同期レプリケーションを使用したハイパーコンバージドに関する考慮事項や導入など、さまざまなシナリオを取り上げます。また、Ethernetに関する一般的な考慮事項、ネットワーク パフォーマンス、動的IPルーティング、ネットワーク仮想化、VMware®環境内での実装、検証方法、監視での推奨事項についても触れています。

対象読者と利用法

本書は、IT管理者、ストレージ アーキテクト、デル・テクノロジーズ™ パートナーと従業員を対象としています。つまり、ネットワーク構築のエキスパートではない方でも理解できるよう内容になっています。ただし、IPネットワーキングを中程度に理解していることを想定しています。

PowerFlex（VxFlex OS）について十分な知識を有している場合は、「PowerFlex機能概要」および「PowerFlexソフトウェア コンポーネント」セクションは飛ばし読みしていただいて構いません。ただし、新しいStorage Data Replicator（SDR）コンポーネントの部分については必ずお読みください。

このガイドでは、ネットワークに関するベスト プラクティスについて最小限知っておくべき事項を説明しています。PowerFlexに適したネットワーク構築のベスト プラクティスや設定について、すべてを詳細に取り上げているわけではありません。PowerFlexのテクニカル エキスパートは、このガイドに記載されている内容よりさらに包括的なベスト プラクティスを推奨する場合があります。

本書の例ではよくCisco Nexus®スイッチが使用されていますが、全般的にどのネットワーク ベンダーにも同じ原則が適用されています。[1]便宜上、通常は、PowerFlexソフトウェア コンポーネントを少なくとも1つ実行しているサーバーを単にPowerFlexノードといい、コンサンプション オプションの区別はありません。

本書全体を通じて**太字**で示す具体的な推奨事項については、この文書の最後にある「推奨事項のサマリー」セクションで再度言及します。

[1] Dell のネットワーク機器の使用に関する一部のガイダンスについては、文書『VxFlex Network Deployment Guide using Dell EMC Networking 25GbE switches and OS10EE』を参照してください。

1 PowerFlex機能概要

PowerFlexはストレージ仮想化ソフトウェアです。サーバー/IPベースのSANを直接アタッチされているストレージから作成することで、柔軟性と拡張性に優れたパフォーマンスと容量をオン デマンドで実現します。従来型SANインフラストラクチャに代わるものとして、PowerFlexではさまざまなストレージ メディアを組み合わせて、ブロック ストレージの仮想プールを作成します。そのパフォーマンス オプションとデータ サービス オプションは一様ではありません。PowerFlexでは、エンタープライズグレードのデータ保護、マルチテナント機能、エンタープライズ機能（インライン圧縮、QoS、シン プロビジョニング、スナップショット、ネイティブ非同期レプリケーションなど）を実現しています。PowerFlexには、次のようなメリットがあります。

大規模な拡張性 – PowerFlexは、1クラスターにノード数個のみから始めて、数百まで拡張することができます。デバイスまたはノードが追加されると、PowerFlexで自動的にデータを均等に再分配して、分散ストレージのプールの完全なバランスを確保します。

超ハイ パフォーマンス – PowerFlexストレージ プールのすべてのストレージ メディア デバイスを使用して、I/O操作を処理します。リソースのこの大規模なI/O並列処理により、ボトルネックが解消されます。スループットとIOPSは、ストレージ プールに追加されたストレージ デバイスの数に正比例して拡張されます。パフォーマンスとデータ保護は、最適化が自動的に行われます。

魅力的な経済性 – PowerFlexには、ファイバー チャネル ファブリックや、HBAのような専用コンポーネントは必要ありません。古くなったハードウェアの大掛かりなアップグレードはありません。障害が発生したか古くなったコンポーネントが単にシステムから除かれる一方、新しいコンポーネントが追加され、データが再バランシングされます。このようにして、PowerFlexでは、ストレージ ソリューションのコストと複雑性が従来型SANよりも軽減されます。

比類のない柔軟性 – PowerFlexは導入オプションに柔軟性があります。2レイヤー導入では、アプリケーションとストレージ ソフトウェアが別々のサーバー プールにインストールされます。2レイヤー導入により、コンピューティング チームとストレージ チームは運用の自立性を維持できます。ハイパーコンバージド導入では、アプリケーションとストレージが1つの共有プールサーバにインストールされ、設置面積とコスト プロファイルを低く抑えます。これらの導入モデルは混在させることもできるので、コンピューティング リソースとストレージ リソースが柔軟性をもって拡張できます。

優れた弾力性 – ストレージ リソースやコンピューティング リソースは、いつでも必要に応じて増減できます。システムでは、データが自動的にオンザフライで再バランシングされます。追加と削除のインクリメントは、小さくしたり大きくしたりできます。容量計画や複雑な再設定は必要ありません。計画外のコンポーネント ロスがあると、再構築動作がトリガーされ、データ保護が保たれます。コンポーネントが追加されると、再バランシングがトリガーされ、利用可能なパフォーマンスと容量が増加します。再構築と再バランシングの動作は、オペレーターが介入しなくてもバックグラウンドで自動的に行われるので、アプリケーションとユーザーのダウンタイムはありません。

エンタープライズ プロバイダーとサービス プロバイダーに不可欠な機能 – サービス品質（QoS）制御により、リソース使用量が動的に管理できるようになり、選択したクライアントで消費できるパフォーマンスの量（IOPSまたは帯域幅）が制限されます。PowerFlexには、データ バックアップとクローニングに向けて、瞬時に書き込み可能なスナップショットがあります。オペレーターは、2つの異なったデータ レイアウトのいずれかを使用してプールを作成することで、ワークロードに最適な環境を確保できます。また、要件が変更されると、ボリュームは異なったプールの間で、ライブ/無停止で移行できます。シン プロビジョニングとインライン データ圧縮により、ストレージの節約と効率的な容量管理が可能になります。またバージョン3.5では、ディザスター リカバリー、データ移行、テスト シナリオ、ワークロードのオフロードに、PowerFlexネイティブの非同期レプリケーションで応じています。

PowerFlexでは、保護ドメインとストレージ プールにより、マルチテナント機能を実現しています。保護ドメインにより、特定のノードやデータセットを分離できます。ストレージ プールは、データ分離、階層化、パフォーマンス管理に使用できます。たとえば、パフォーマンスへの要求が高いビジネス クリティカルなアプリケーションとデータベースのデータは、ハイパフォーマンスのSSD、NVMe、またはSCMベースのストレージ プールに最も低レイテンシーで保存され、アクセス頻度の低いデータは、低コストで大容量のSSDから低めのDWPD仕様で構築されたプールに保存されます。繰り返しますが、1つのボリュームから別のボリュームへはライブで、ワークロードを中断することなしに移行されます。

2 PowerFlexソフトウェア コンポーネント

PowerFlexは基本的に3つのタイプのソフトウェア コンポーネント、Storage Data Server（SDS）、Storage Data Client（SDC）、Meta Data Manager（MDM）で構成されています。バージョン3.5ではレプリケーションを可能にする新しいコンポーネント、Storage Data Replicator（SDR）が導入されます。

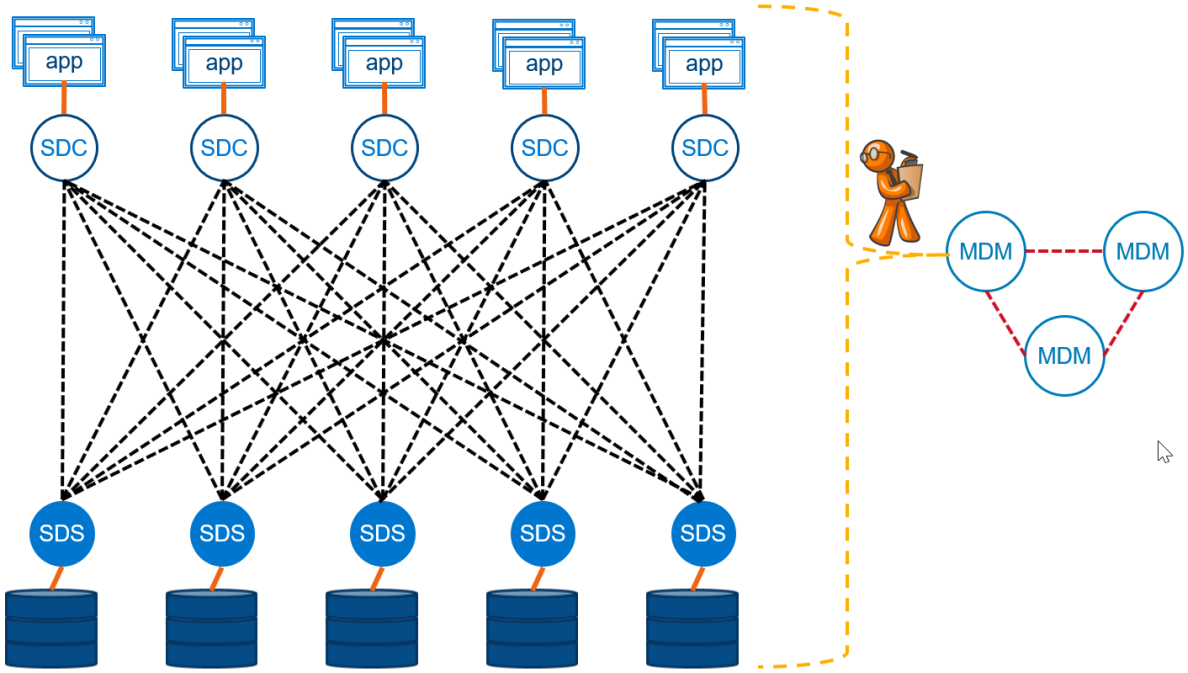

図 1 PowerFlex導入の論理的な例。SDCで使用可能な各ボリュームは、SDSを実行する多数のシステムにわたって分散されており、各SDCには、ボリュームにサービスを提供する各SDSへの冗長パスがあります。Meta Data Manager（MDM）クラスターはデータ パスの外部に存在しており、システムの監視、データ レイアウトの調整、変更が発生した場合のSDCの更新を行います。

2.1 Storage Data Server（SDS）

Storage Data Server（SDS）は、ローカルのrawストレージをノードにアグリゲートし、PowerFlexクラスターの一部として分配するユーザー スペース サービスです。SDSは、サーバー側のソフトウェア コンポーネントです。データを他のノードへ分配する役割を担うサーバーには、SDSサービスがインストールされ、そこで実行されます。SDSのコレクションは、PowerFlex永続レイヤーを形成します。

複数のSDSがともに動作して、ユーザー データの冗長コピーを保持し、相互にハードウェア ロスから保護し、ハードウェア コンポーネントに障害が発生したときはデータ保護を再構築します。SDSで活用されるのは、SSD、PCIeベースのフラッシュ、ストレージ クラス メモリー、回転ディスク メディア、利用可能なRAM、またはこれらの任意の組み合わせです。

SDSは、さまざまなフレーバーのLinuxや、ESXi上の仮想アプライアンスで、ネイティブに実行される場合があります。PowerFlexクラスターのSDSの最大数は512になります。

SDSコンポーネントは相互に直接通信することができ、SDSのコレクションは完全にメッシュ化されています。SDSは、再構築、再バランシング、I/O並列処理に向けて最適化されています。ユーザー データ レイアウトは、SDSコンポーネント間で、**ストレージ プール、保護ドメイン、フォールト セット**によって管理されます。

SDCで使用されるクライアント ボリュームは、**ストレージ プール**の内部に配置されます。ストレージ プールは、同じようなタイプのストレージ メディアをドライブレベル単位で論理的にアグリゲートするために使用されます。ストレージ プールで提供されるストレージ サービスは、容量とパフォーマンスによってさまざまなレベルに区別されています。

ノード、デバイス、ネットワーク接続の障害からの保護は、ノードレベル単位で**保護ドメイン**を通じて管理されます。保護ドメインとは、ユーザー データ レプリカが保持されているSDSのグループです。

フォールト セットでは、障害が同時に発生しやすいノード セット（ラック全体など）に冗長コピーが存在しないようにすることで、非常に大規模なシステムの複数のノードで同時に障害が発生しても許容されます。

2.2 Storage Data Client（SDC）

Storage Data Client（SDC）により、オペレーティング システムやハイパーバイザーでは、PowerFlexクラスターによって分配されるデータにアクセスできるようになります。SDCはクライアント側ソフトウェア コンポーネントであり、Windows®、さまざまなフレーバーのLinux、IBM AIX®、ESXi®その他でネイティブに実行されますソフトウェアHBAに類似していますが、複数のネットワーク パスやエンドポイントを並列で使用するよう最適化されています。

SDCは、それを実行するオペレーティング システムまたはハイパーバイザーに、「ボリューム」と呼ばれる論理ブロック デバイスへのアクセスを提供します。ボリュームは、従来型SANのLUNに類似しています。各論理ブロック デバイスは、データベースまたはファイル システムのロー ストレージとなり、クライアント ノードにはローカル デバイスとして出現します。

SDCでは、ボリュームでのブロックの場所に基づき、どのStorage Data Server（SDS）エンドポイントにコンタクトするかを把握しています。SDCでは、分散ストレージ リソースを、PowerFlexを実行している他のシステムから直接消費します。SDC間で、1つのプロトコル ターゲットやネットワーク エンドポイントを他のSDCと共有することはありません。SDC間で、負荷は均等かつ自律的に分配されます。

SDCは非常に軽量です。SDCからSDSへの通信は、ストレージ プールに寄与するSDSストレージ サーバーすべてにわたって、本質的にマルチパス化されています。これは、iSCSIのように複数のクライアントで1つのプロトコル エンドポイントをターゲットにするアプローチとは対照的です。広く分散しているというSDC通信の特性により、パフォーマンスと拡張性が大幅に向上します。

SDCにより、クラスタリングのような使用に対する共有ボリューム アクセスが可能になります。SDCには、iSCSIイニシエーターも、ファイバー チャネル イニシエーターも、FCoEイニシエーターも必要ありません。SDCは、シンプル性、高速性、効率性を追求して最適化されています。PowerFlexクラスターのSDCの最大数は1024になります。

2.3 Meta Data Manager（MDM）

MDMでは、PowerFlexシステムの動作を制御します。クライアントとそのボリューム データとの間のマッピングを決定してパブリッシュし、システムの状態を追跡し、また、再構築や再バランシングのディレクティブをSDSコンポーネントへ発行します。

MDMは、PowerFlexでquorumのノーションを確立します。PowerFlexで唯一、タイトにクラスター化されているコンポーネントです。影響力があり、冗長化されており、可用性に優れています。再構築や再バランシングのようなI/O操作時やSDS to SDS操作時に調べられることはありません。もっとも、ハードウェア コンポーネントに障害が発生した場合には、MDMクラスターから自動ヒーリング操作を数秒以内に開始するよう指示されます。MDMクラスターは少なくともサーバー3つで構成され、これでquorumは維持されますが、可用性を向上させるには5つ使用します。MDMクラスターは3ノードまたは5ノードで、いずれの場合もプライマリーは常に1つです。セカンダリーMDMが1つまたは2つ、タイ ブレーカーが1つまたは2つとなります。

2.4 Storage Data Replicator（SDR）

バージョン3.5以降では、PowerFlexクラスター間の非同期レプリケーションを促進するという、新しいオプションのソフトウェアが導入されています。Storage Data Replicator（SDR）は、レプリケーションが採用されていない場合、通常のPowerFlex操作では必要ありません。ソース側で、SDRはSDCとSDSとの間に仲介者として立ち、ボリュームのアドレス空間の関連する部分をホストします。ボリュームがレプリケートされる際、SDCはSDRへ書き込みを送信し、そこで書き込みが分割され、両方ともレプリケーション ジャーナルに書き込まれ、関連するSDSサービスへ転送され、ローカル ディスクへ引き渡されます。

SDRは、MDMからクローズするインターバルについての指示があるまで、書き込みをインターバルジャーナルに蓄積します。ボリュームがマルチボリューム レプリケーション コンシステンシー グループの一部である場合は、インターバル クロージャーが複数同時に発生します。書き込みフォールディングが適用され、インターバルは転送キューへ追加されてターゲット側へ伝送されす。

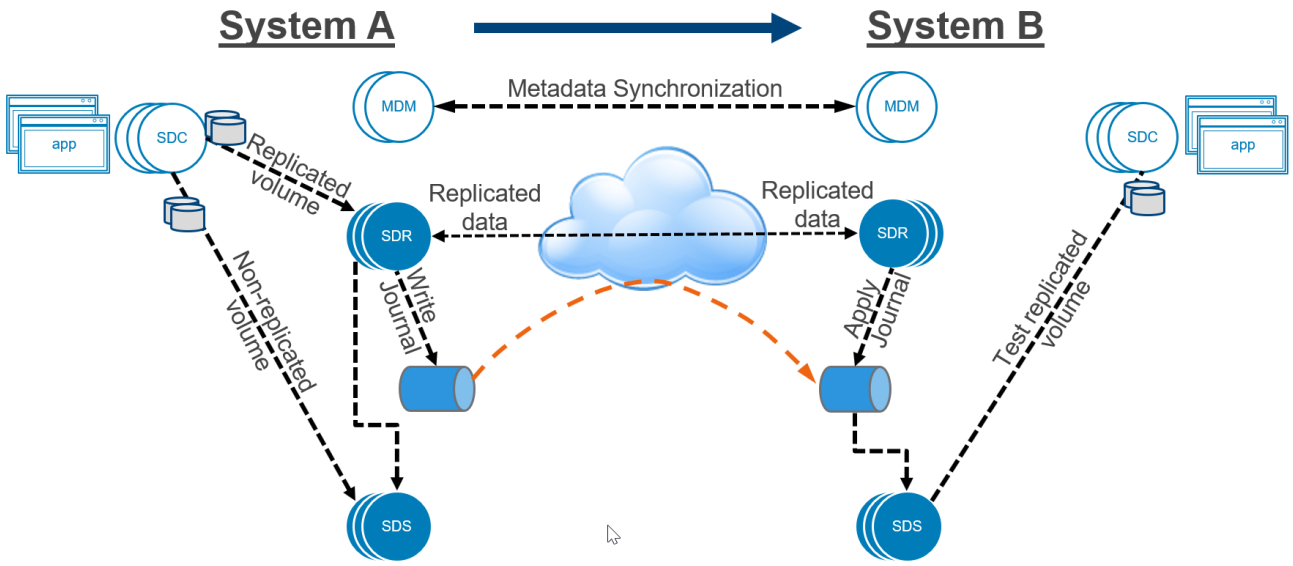

図 2　　レプリケーション データ フローをシンプル化した図。

ターゲット側のSDRでは、データを別のジャーナルで受信してSDSへ送信し、ターゲットのレプリカ ボリュームに応用します。

3 トラフィック タイプ

PowerFlexのパフォーマンス、拡張性、セキュリティがメリットとなるのは、ネットワーク アーキテクチャにPowerFlexトラフィック パターンが反映されている場合です。これは特に、PowerFlexの大規模な導入に当てはまります。PowerFlexを構成するソフトウェア コンポーネント（SDC、SDS、MDM、SDR）は、予測可能な方法で相互に対話します。**PowerFlexの導入を設計するアーキテクトは、ネットワーク レイアウトについて情報に基づく選択をするために、こうしたトラフィック パターンを認識しておく必要があります。**

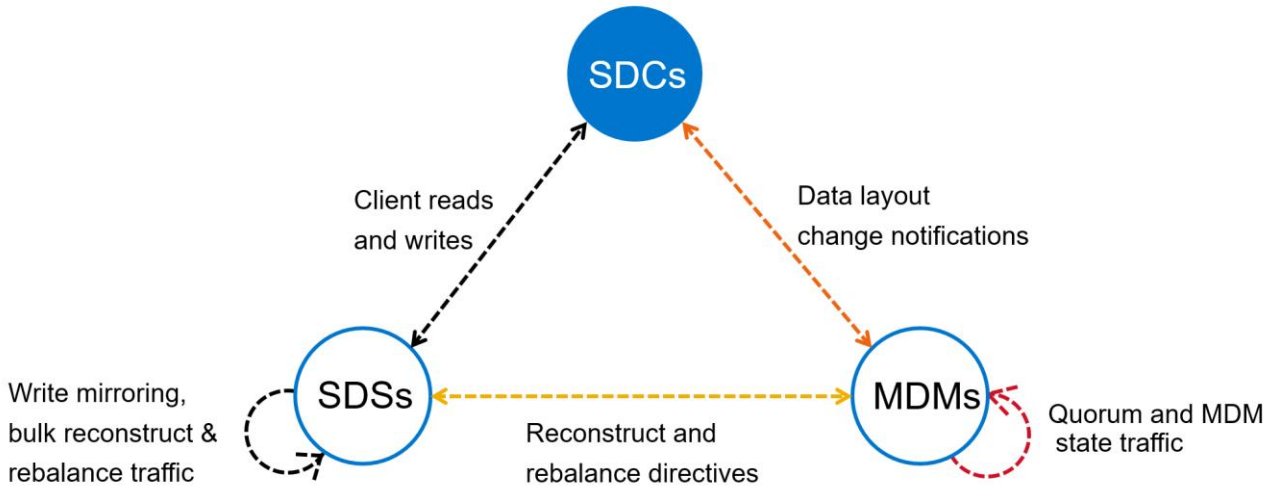

図 3　PowerFlex基本ソフトウェア コンポーネントでの通信方法をシンプル化した図。PowerFlexシステムは、多くのSDC、SDS、MDMを備えています。この図では、SDC、SDS、MDMをグループ化しています。SDSやMDMから自身へ戻るように指している矢印は、他のSDSやMDMとの通信を表現したものです。SDCからSDCへの通信は存在しないことに注意してください。トラフィック パターンは、SDC、SDS、MDMの物理的な場所に関係なく同一です。

次の説明では、フロントエンド トラフィックとバックエンド トラフィックとを区別しています。これは論理的な区別であり、ネットワークを物理的に区別する必要はありません。PowerFlexでは、フロントエンドとバックエンドの両方のトラフィックを同じ物理ネットワーク上で実行することも、別々のネットワークに分離することもできます。必須ではありませんが、ストレージ ネットワークでは多くの場合、フロントエンドとバックエンドのトラフィックを独立させる方が好ましいです。

たとえば、このような分離は、インフラストラクチャの別個の部分をそれぞれ別のチームで管理するという、運用上の理由によって行われる場合があります。ただし、バックエンド トラフィックを分離する理由として最もよくあるのは、再構築と再バランシングのパフォーマンスを向上させることができる、というものです。また、フロントエンド トラフィックを分離することで、ネットワーク上での競合が回避され、再構築/再バランシング動作中のクライアントやアプリケーション トラフィックに対するレイテンシーの影響が軽減されます。

3.1 Storage Data Client（SDC） to Storage Data Server（SDS）

SDCとSDSとの間のトラフィックは、フロントエンド ストレージ トラフィックの大部分を形成します。フロントエンド ストレージ トラフィックには、クライアントで受信したりクライアントから送信したりする読み取り/書き込みトラフィックがすべて含まれています。このネットワークはスループット要件が高度です。

3.2 Storage Data Server（SDS）からStorage Data Server（SDS）

SDS間のトラフィックは、バックエンド ストレージ トラフィックの大部分を形成します。バックエンド ストレージ トラフィックには、SDS間でミラーリングされた書き込み、トラフィックの再バランシング、トラフィックの再構築、ボリューム移行トラフィックが含まれます。このネットワークはスループット要件が高度です。

3.3 Meta Data Manager（MDM） to Meta Data Manager（MDM）

MDMは、クラスター内部で動作を調整するために使用されます。PowerFlexに対して、トラフィックを再バランシング、再構築、リダイレクトするためのディレクティブを発行します。また、レプリケーション コンシステンシー グループを調整し、レプリケーション ジャーナル インターバル クロージャーを決定し、PowerFlexレプリカピア システムとのメタデータ同期を保持します。MDMは冗長化されており、また、相互の通信を途切れさせないようにして、quorumを確立し、データ レイアウトについての把握の共有を維持する必要があります。

MDMでは、I/Oトラフィックを送信したり、I/Oトラフィックに直接干渉したりすることはありません。MDM間で交換されるデータは比較的軽量であり、SDSやSDCのトラフィックに要するのと同じレベルのスループットは要求されません。ただし、MDMでは、100msに1回発生するquorum交換は、タイムアウトが非常に短いです（< 400ms）。**MDM to MDMトラフィックには、安定していて信頼できる低レイテンシーのネットワークが必要です。**MDM to MDMトラフィックは、バックエンド ストレージ トラフィックと見なされます。PowerFlexは、MDM間のトラフィック専用のネットワークを1つ以上使用することをサポートしています。本番稼働環境では、少なくとも10 GbEリンクをMDMごとに2つ使用する必要がありますが、25 GbEの方が一般的です。

PowerFlex 3.5では、レプリケーション ピア システム間でのクロスクラスターMDM to MDMトラフィックを導入しています。ここでのMDMによる通信では、レプリケーション フローとジャーナルの状態を制御する必要があります。統合レプリケーションの状態は、ソース サイトとデスティネーション サイトとの間で同期されます。MDM to MDMピア メタデータの同期は、WANで発生し、レイテンシー200ms未満で行われる必要があります。

3.4 Meta Data Manager（MDM） to Storage Data Client（SDC）

プライマリー（ソフトウェアがマスターを呼び出すもの）MDMは、データ レイアウトが変更された場合に、SDCと通信する必要があります。これは、SDCでSDCのボリューム ストレージをホストするSDSが、追加された、削除された、メンテナンス モードになった、またはオフラインになったことが原因で発生します。ボリュームがレプリケーション コンシステンシー グループに配置された場合に発生することもあります。プライマリーMDMとSDCとの間の通信は遅延していて非同期であっても、信頼できる低レイテンシーのネットワークを必要としています。MDM to SDCトラフィックは、フロントエンド ストレージ トラフィックと見なされます。

3.5 Meta Data Manager（MDM） to Storage Data Server（SDS）

プライマリーMDMは、SDSとデバイスの正常性を監視し、再バランシング/再構築ディレクティブを発行するため、SDSと通信する必要があります。MDM to SDSトラフィックには、信頼できる低レイテンシーのネットワークが必要です。MDM to SDSトラフィックは、バックエンド ストレージ トラフィックと見なされます。

3.6 Storage Data Client（SDC） to Storage Data Replicator（SDR）

ボリュームがレプリケートされている場合は、通常のSDC to SDSトラフィックがSDR経由でルーティングされます。ボリュームがレプリケーション コンシステンシー グループに配置されている場合、MDMでは、SDCに提示されているボリューム マッピングを調整し、SDCに指示してI/O操作をSDRへ発行し、そこから関連SDSへ渡されるようにします。そのSDRは、SDCには単なる別のSDSであるかのように出現します、SDC to SDRトラフィックはスループット要件が高度なので、信頼できる低レイテンシーのネットワークが必要です。SDC to SDRトラフィックは、フロントエンド ストレージ トラフィックと見なされます。

3.7 Storage Data Replicator（SDR） to Storage Data Server（SDS）

ボリュームがレプリケートされ、I/OがSDCからSDRへ送信されると、ソース システムにはSDRからSDSへの後続I/Oが2つ発生します。まず、SDRから関連するSDSへボリュームI/Oを移して処理（圧縮など）し、ディスクへ引き渡します。次に、SDRは書き込みをジャーナリング ボリュームへ適用します。ジャーナル ボリュームはPowerFlexシステムでは単に別のボリュームであるため、SDRではI/Oを、ジャーナル ボリュームが存在するストレージ プールを構成するディスクのSDSへ送信しています。

ターゲット システムのSDRでは、受信した一貫性のあるジャーナルを、レプリカ ボリュームをバックアップするSDSに適用します。これらの各ケースでのSDRは、あたかもSDCであるかのように動作します。それでもかかわらず、SDR to SDSトラフィックは、バックエンド ストレージ トラフィックと見なされます。SDR to SDSトラフィックは高スループットになり、レプリケートされるボリュームの数に比例します。信頼できる低レイテンシーのネットワークが必要です。

3.8 Meta Data Manager（MDM） to Storage Data Replicator（SDR）

MDMは、ジャーナルインターバル クロージャーを発行し、RPOコンプライアンスを収集して報告し、デスティネーション ボリュームで一貫性を維持するため、SDRと通信する必要があります。ピア システムから転送されるレプリケーション状態を使用して、MDMはそのローカルSDRに命令してジャーナル操作を実行します。

3.9 Storage Data Replicator（SDR） to Storage Data Replicator（SDR）

ソース内のSDRやターゲットPowerFlexクラスター内のSDRは、相互の通信をしません。しかし、ソース システムのSDRは、レプリカ ターゲット システムのSDRと通信します。SDRは、LANまたはWANのネットワークを介して、ジャーナル インターバルをデスティネーションSDRへ送出します。SDR　SDRトラフィックは、レイテンシーにさほど敏感ではありませんが、ラウンドトリップ時間が200msを超えないようにする必要があります。

3.10 他のトラフィック

PowerFlexクラスターには、他にもさまざまなタイプの低ボリューム トラフィックがあります。他のトラフィックには、低頻度の管理、インストール、レポート作成などがあります。これには、PowerFlex Gateway（REST API Gateway、Installation Manager、SNMPトラップ センダー）へのトラフィック、vSphereプラグイン、PowerFlex Manager、Light Installation Agent（LIA）との間でのトラフィック、MDMへのレポート作成トラフィックまたは管理トラフィック（レポート作成用syslogや管理者認証用LDAPなど）も含まれます。また、MDM、SDS、SDCの間でのCHAP認証トラフィックも含まれます。詳細については、PowerFlex Technical Resource Centerの『Getting to Know Dell EMC PowerFlex』ガイドを参照してください。

SDCは、他のSDCとは通信しません。これは、プライベートVLANとネットワーク ファイアウォールを使用して強制されます。

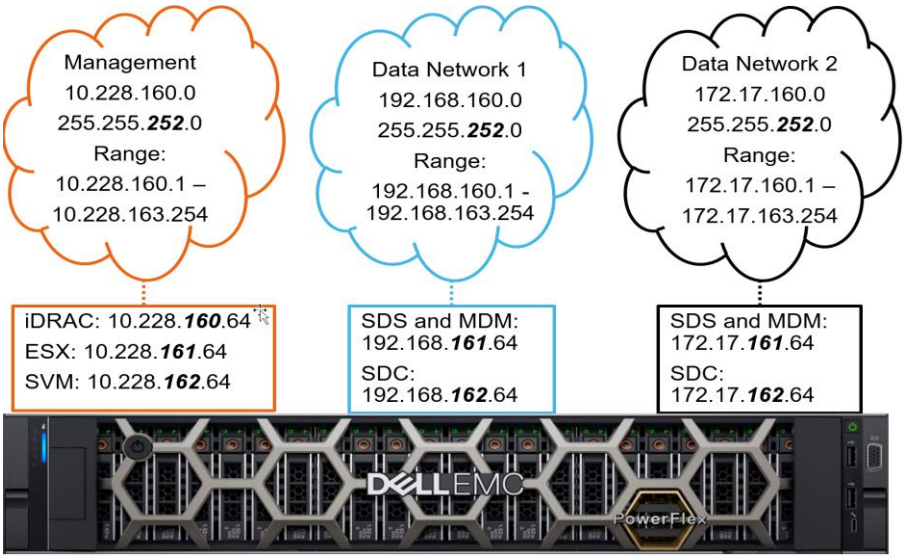

図 4　シンプルなPowerFlexハイパーコンバージド ネットワーク レイアウト。管理ネットワークはルーティングされ、iDRAC、ESX、ストレージ仮想マシン（SVM）へのアクセスを提供します。冗長ネットワークでは、SDS、MDM、SDCトラフィックを伝送します。SDSトラフィックとMDMトラフィックとでは、同じIPアドレス セットを使用します。このトラフィックは、大規模導入の場合と同様、フロントエンド トラフィック（SDS、SDC、MDM）やバックエンド トラフィック（SDS、MDM）へのセグメント化はされません。MDM仮想IPに使用できるアドレス空間は、192.168.160.Xと172.17.160.Xです。

4 PowerFlex TCPポートの使用方法

PowerFlexは、Ethernetファブリックを介して動作します。多くのPowerFlexプロトコルは独自仕様ですが、すべての通信に標準TCP/IPトランスポートが使用されています。

次の図に、PowerFlexソフトウェア コンポーネント間でのポートの使用と通信の大まかな概要を示します。一部のポートは固定されており変更もされませんが、それ以外のポートは設定可能であり別のポートへ再割り当てされることもあります。リスト全体とカテゴリー分けについては、『Dell EMC PowerFlex Security Configuration Guide』の「Port usage and change default ports」セクションを参照してください。

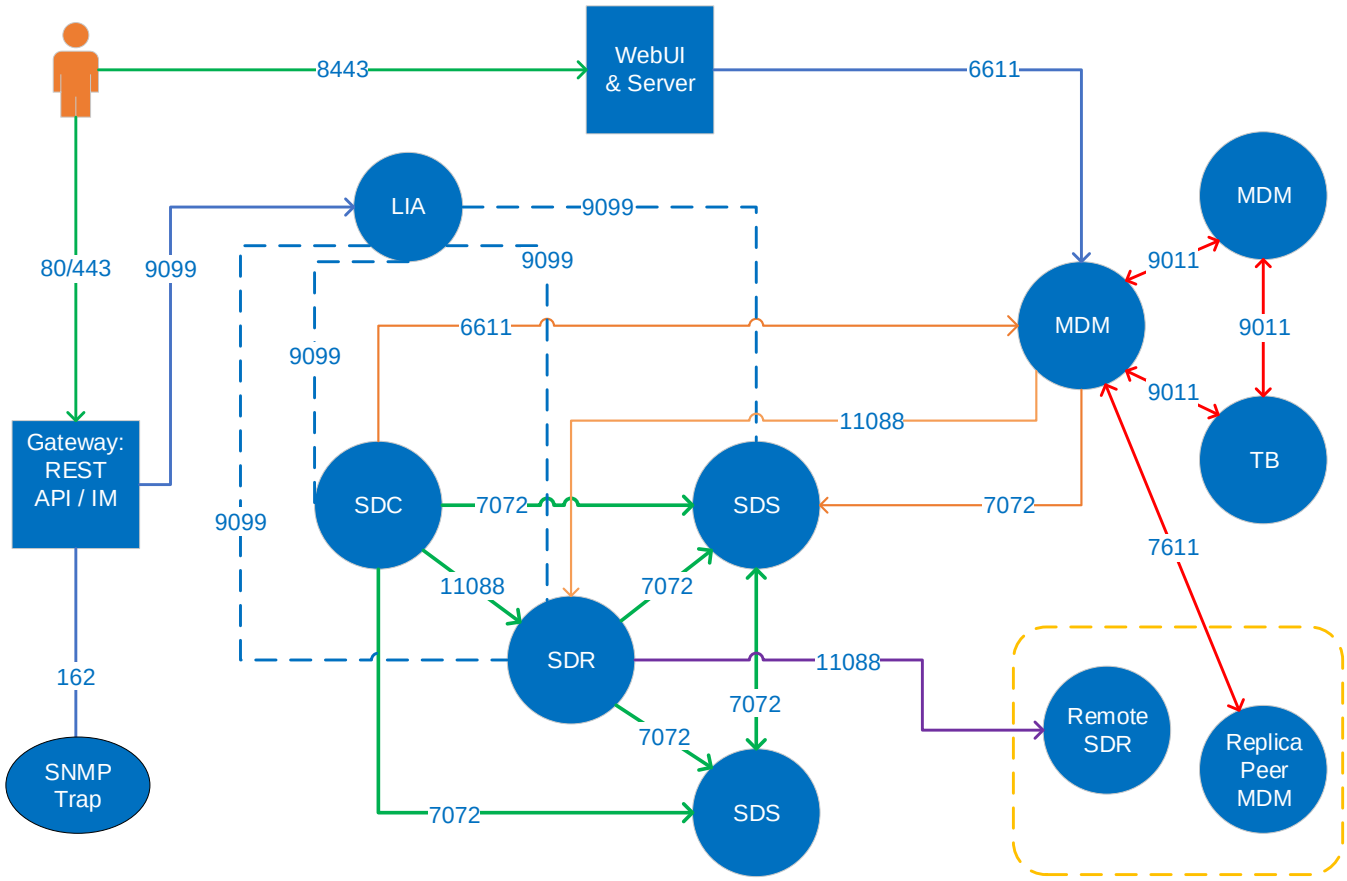

図 5 PowerFlexソフトウェアデファインド ストレージ コンポーネント内でのTCPポートの使用と通信。図の矢印は、接続の開始方向を示します。つまり、矢印はリスニング サービス ポートを指しているのです。接続の開始後、データは双方向に移動できます。破線は、通信がノードの内部、設置されているコンポーネント間で行われることを示します。

ポート（MDMでは25620と25600、SDSでは25640）もリッスンしていることがあります。これらはPowerFlex内部デバッグツールによってのみ使用され、日常的な運用とトラフィックの一部ではありません。

5 ネットワーク フォールト トレランス

PowerFlexコンポーネント（MDM、SDS、SDC、SDR）の間での通信は、異なった物理ネットワーク上の、少なくとも2つのサブネットへ割り当てる必要があります。PowerFlexのネットワーキング層は、これらの各コンポーネントで、割り当てられている多数のサブネットにわたって、ネイティブ リンクのフォールト トレランスとマルチパス化を実現します。この結果、設計上は次のような利点があります。

1. リンク障害が発生した場合、PowerFlexはほぼ即座に問題を認識し、帯域幅のロスを調整します。
2. スイッチベースのリンク アグリゲーションが使用されていた場合、PowerFlexには1つのリンク ロスを識別する手段はありません。
3. PowerFlexは、リンクに障害が発生した場合に、MDM/SDS/SDCコンポーネントへ割り当てられているサブネット全体にわたって、通信を2～3秒以内で動的に調整します。これは、SDS　SDS接続とSDC　SDS接続で特に重要です。
4. これらの各コンポーネントには、最大8つのサブネットにわたるロード バランシングをしてトラフィックをアグリゲートする機能があり、スイッチベースのリンク アグリゲーションのメンテナンスでの複雑さが軽減されます。また、ストレージ レイヤー自体で管理しているため、スイッチベースのアグリゲーションよりも効率が高く、メンテナンスがシンプルになります。

   注意：以前のバージョンのPowerFlexソフトウェアでは、リンク関連の障害が発生した場合、SDC→SDSネットワークでは最大17秒のネットワーク サービスの中断とI/O遅延が発生する可能性がありました。SDCには通常15秒のタイムアウトがあり、I/Oがが再発行されるのは、タイムアウトに達したときにデッド ソケットがすでにクローズしている場合のみであり、別の「good」ソケットで行われます。

   バージョン3.5以降のPowerFlexは、I/Oタイムアウトへの依存はしなくなっていますが、リンク切断通知は使用します。リンク ダウン イベントが発生すると、関連するTCP接続はすべて2秒後に閉じられ、応答を受信していないインフライトI/Oメッセージはすべてアボートされ、I/OがSDCによって再発行されます。

ネイティブ ネットワーク パスのロード バランシングとスイッチベースのリンク アグリゲーションの両方が完全にサポートされていますが、多くの場合、ネイティブ ネットワーク パスのロード バランシングに依存する方がシンプルです。必要に応じて、こうしたアプローチを組み合わせることで、たとえば、データパス ネットワークを2つ、各論理ネットワークでノードあたり2つの物理ポートを使用するトランクを介して作成することができます。

PowerFlex Managerでは、まさにこのことを、アプライアンスに対して実行します。リンク アグリゲーションとネイティブのマルチパス化とを組み合わせて使用することで、ネットワーク フォールト トレランスをレイヤー化し堅牢なものにします。『Dell EMC PowerFlex Appliance Network Planning Guide』を参照してください。

6 ネットワーク インフラストラクチャ

現在、PowerFlex で最も一般的に使用されているトポロジーは、リーフ/スパインとフラット ネットワークです。フラット ネットワークは、小規模なネットワークで使用されています。最新のデータセンターでは、レガシーの階層型トポロジーよりもリーフ/スパイン トポロジーが好ましいとされています。このセクションでは、PowerFlexデータ トラフィック用のトランスポート メディアとして、フラットとリーフ/スパインとでトポロジーを比較します。

デル・テクノロジーズでは、ノンブロッキング ネットワーク設計の使用が推奨されます。ノンブロッキング ネットワーク設計により、メッセージ ループを回避するために一部のネットワーク ポートをブロックすることなしに、すべてのスイッチ ポートを同時に使用できるようになります。したがって、デル・テクノロジーズでは、PowerFlexをホストしているネットワーク上でスパニング ツリー プロトコル（STP）を使用することに対し、強く推奨します。パフォーマンスを最大化しサービス品質（QoS）を予測可能にするため、ネットワークのオーバーサブスクライブはしないでください。

6.1 リーフ/スパイン ネットワーク トポロジー

2層のリーフ/スパイン トポロジーには、リーフ スイッチ間にスイッチ ホップが1つあり、エンドポイント間には帯域幅を大きく取っています。リーフ/スパイント ポロジーを適切にサイジングすると、アップリンク ポートのオーバーサブスクリプションが解消されます。非常に大規模なデータセンターでは、3層のリーフ/スパイン トポロジーを使用することもあります。シンプルにするため、本書では2層のリーフ/スパインの導入に焦点を当てます。

リーフ/スパイン トポロジーでは、どのリーフ スイッチもすべてのスパイン スイッチに接続されています。リーフ スイッチは、他のリーフ スイッチへ直接接続する必要はありません。スパイン スイッチは、他のスパイン スイッチへ直接接続する必要はありません。

ほとんどの場合、デル・テクノロジーズでは、リーフ/スパイン ネットワーク トポロジーを使用することが推奨されます。 これは次の理由からです。

- PowerFlexは、1つのクラスターで何百ものノードへスケールアウトできます。
- リーフ/スパイン アーキテクチャは将来を見据えたものです。ネットワークのアーキテクチャを作り直さなくとも、スケールアウト導入を容易にすることができます。
- リーフ/スパイン トポロジーでは、すべてのネットワーク リンクを同時に使用することができます。レガシーの階層型トポロジーでは、スパニング ツリー プロトコル（STP）などのテクノロジーを採用して、ループを回避するために一部のポートをブロックする必要があります。
- リーフ/スパイント ポロジーを適切にサイジングすると、アップリンク オーバーサブスクリプションの解消により、レイテンシーがさらに予測可能になります。

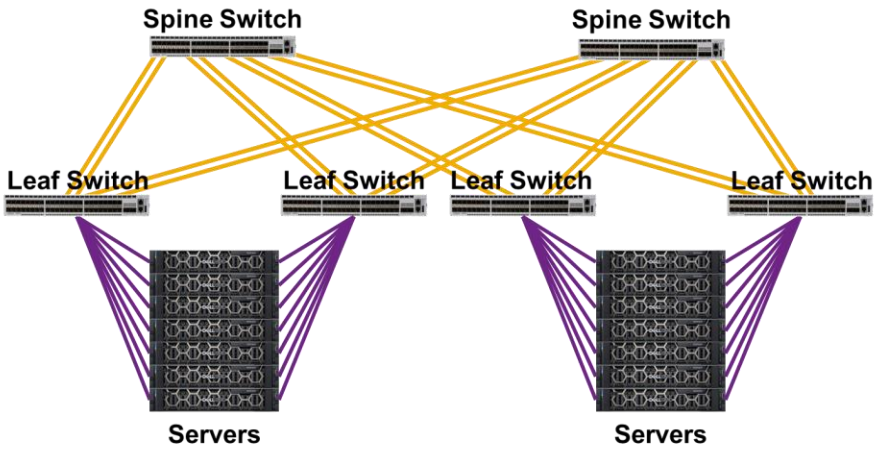

図 6　2層のリーフ/スパイン ネットワーク トポロジー。各リーフ スイッチから他のいずれのリーフ スイッチへも、複数のパスがあります。すべてのリンクがアクティブです。これにより、ネットワーク上のデバイス間でのスループットが向上します。リーフ スイッチは、MLAG（ここには示しません）で使用するため相互に接続されている場合があります。

6.2　フラット ネットワーク トポロジー

フラット ネットワーク トポロジーは、既存のフラット ネットワークが拡張されている場合や、ネットワークの拡張が想定されていない場合には、より実装しやすく好ましいものとなります。フラット ネットワークでは、すべてのスイッチがホストを接続するために使用されます。スパイン スイッチはありません。

ただし、少数のアクセス スイッチよりも拡張する場合は、追加のクロスリンク ポートが必要になり、フラット ネットワーク トポロジーのコストが法外な額になる可能性があります。フラット ネットワーク トポロジーのユース ケースは、概念実証の導入や小規模データセンターの導入などであり、ラックの数が数個から増加することはありません。

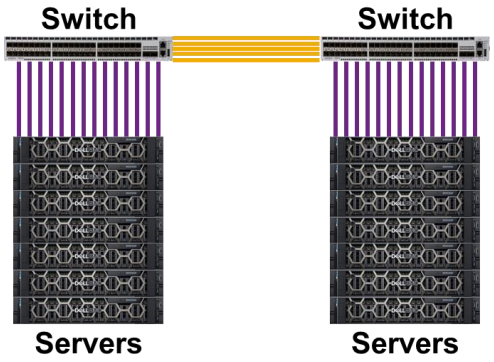

図 7　フラット ネットワーク。このネットワーク設計では、コストと複雑さは軽減されますが、冗長性と拡張性は犠牲になります。このビジュアルでは、どのスイッチも単一障害点です。MLAG（ここには示しません）などのテクノロジーを使用すると、単一障害点を発生させることなしに、フラット ネットワークを構築することができます。

7 ネットワーク パフォーマンスとサイジング

ネットワークを適切にサイジングすると、ネットワーク管理者やストレージ管理者は、個々のポートやリンクがパフォーマンスや運用上のボトルネックとなる懸念から解放されます。エンドポイントのホットスポットではなくネットワークを管理することは、PowerFlexのアーキテクチャで重要な利点です。

PowerFlexでは、I/Oがネットワークの複数のポイントにわたって均等に分散されるため、ネットワーク パフォーマンスは適切にサイジングする必要があります。

7.1 ネットワーク レイテンシー

ネットワークを設計する際、ネットワーク レイテンシーを考慮することは重要です。 ネットワーク レイテンシーを最短にすると、パフォーマンスと信頼性を向上させることができます。**ベスト パフォーマンスを得るには、すべてのSDSやSDCの通信のレイテンシーは、通常の動作条件でのネットワークのみのラウンドトリップ時間1ミリ秒を超えないようにする必要があります。**ワイド エリア ネットワーク（WAN）の最短レスポンス タイムは通常、この制限を超えているため、PowerFlexクラスターはWANで運用しないでください。

非同期レプリケーションを実装しているシステムも、通常の通信やSDC、MDM、SDS通信に関しては、この例外ではありません。データは独立したPowerFlexクラスター間でレプリケートされ、どのクラスターもそれ自体で1ミリ秒以下ルールに従っていることが必要です。その違いは、ピア化されているシステムの間でのレイテンシーです。非同期レプリケーションは通常WANで行われるため、レイテンシー要件がさほど制限的なものでないことは必然です。**ただし、PowerFlexクラスター コンポーネント間でのネットワーク レイテンシーは、MDM←→MDMであってもSDR←→SDRであっても、200msのラウンド トリップ時間を超えないようにする必要があります。**

レイテンシーは、すべてのコンポーネント間で、双方向でテストする必要があります。これは、pingを行うことによって、より広範囲にはSDS Network Latency Meter Testによって検証できます。オープンソース ツールiPerfを使用すると、帯域幅を検証できます。iPerfはデル・テクノロジーズではサポートされていませんので、ご注意ください。PowerFlexの導入を検証するために使用されるiPerfなどのツールについては、本書の「検証方法」セクションで詳しく取り上げます。

7.2 ネットワーク スループット

PowerFlex実装を設計する際には、ネットワーク スループットが重要なコンポーネントとなります。障害が発生したノードの再構築にかかる時間を短縮すること、データ分散が不均一になった場合にデータを再配布するのにかかる時間を短縮すること、ノードで提供できるI/Oの量を最適化すること、パフォーマンスの期待事項に応えることには、スループットが重要です。

PowerFlexソフトウェアは、テスト目的や調査目的で1Gbネットワークに導入することができますが、ストレージ性能がネットワーク容量によってボトルネック化する可能性があります。**最低でも、10Gbのネットワーク テクノロジーを、25Gbテクノロジーを推奨最小リンク スループットとして活用することが推奨されます。**現在のすべてのPowerFlexノードは、ポートが少なくとも4つ、それぞれ最小ポート帯域幅を25 GbEにして出荷されており、将来を見据えたオプションとして100 GbEポートが付随しています。これは、レプリケーションのケースと帯域幅要件の追加を考慮する場合、特に重要です。

さらに、PowerFlexクラスター自体は異種混在型である場合もありますが、**保護ドメインを構成するSDSコンポーネントは、同等のストレージとネットワーク パフォーマンスを備えたハードウェア上に配置する必要があります。**これは、I/O時や、寄与するコンポーネントすべてにわたってボリューム データのストライピングが広範に行われることによる再構築/再バランシング動作時に、保護ドメインの総帯域幅が最も脆弱なリンクによって制限されるためです。最も遅いメンバーよりも速く移動することのないハイキング パーティーのようなものだと考えてください。

異種混在型OSとハイパーバイザーとが混在した組み合わせをする場合にも、同様の考慮事項があります。VMwareベースのハイパーコンバージド インフラストラクチャは、仮想化オーバーヘッドがあるため、ベアメタル構成よりもパフォーマンスが低くなります。また、HCIとベアメタルのノードを保護ドメインで混在させると、最も遅いメンバーのパフォーマンス機能も含むストレージ プールのスループットは制限されます。（ストレージ ソフトウェアから見ると）可能性があり許容されますが、ユーザーはこの含意に留意する必要があります。PowerFlexラックまたはアプライアンスでサポートされる構成ではありません。

スループットに関する考慮事項に加えて、**各ノードには、スループット要件に関係なく、少なくとも2つの別個のネットワーク接続で冗長性を持たせることが推奨されます。**これは、ネットワーク テクノロジーが向上しても変わらず重要です。たとえば、40Gbリンク2つを100Gbリンク1つに置き換えると、スループットは向上しますが、リンクレベルのネットワーク冗長性が犠牲になります。

ほとんどの場合、ノードのネットワーク スループットの量は、ノード上でホストされているストレージ メディアを組み合わせた最大スループットと一致しているか、上回っている必要があります。言い換えれば、ノードのネットワーク要件は、基盤となるストレージ メディアを総合したパフォーマンスに比例しています。

必要となるネットワーク スループットの量を判断する場合、最新のメディア パフォーマンスは通常1秒あたりのメガバイト単位で測定されるが、最新のネットワーク リンクは通常1秒あたりのGb単位で測定されることに注意してください。

1秒あたりのメガバイトを1秒あたりのGbに変換するには、まずメガバイトに8を乗算してメガビットに変換してからメガビットを1000で除算すると、Gbになります。

$$\text{gigabits} = \frac{\text{megabytes} * 8}{1{,}000}$$

これは、PowerFlexの標準である二進法の「キロ」の定義1024については考慮していないため、完全に正確ではないことに注意してください。ただし、本書での説明目的にはこれで十分です。

7.2.1 例：SDSのみ（ストレージのみ）のノードでSSDは10個

SDSのみをホストしている1Uノードを前提としています。これはハイパーコンバージド環境ではないため、考慮する必要があるのはストレージ トラフィックのみです。ノードにはSAS SSDドライブが10個搭載されています。これらのドライブはそれぞれ、最適な状況（シーケンシャルI/O、PowerFlexは再構築/再バランシング動作時に最適化される）で、rawスループットが1秒あたり1000メガバイトになります。したがって、基盤となるストレージ メディアの総スループットは1秒あたり10,000メガバイトです。

$$10 * 1000 \text{ megabytes} = 10{,}000 \text{ megabytes}$$

次に、先に説明した数式を使用して、10,000メガバイトをGbに変換します。まず10,000 MBを8で乗算し、それから1,000で除算します。

$$\frac{10{,}000 \text{ megabytes} * 8}{1{,}000} = 80 \text{ gigabits}$$

この場合、ノード上のすべてのドライブが読み取り操作サービスを可能な限り最大速度で行っている場合、ネットワークに必要な総スループットは1秒あたり80ギガになります。当社では読み取り操作のみを対象としています。これは通常の場合、ネットワーク帯域幅要件を見積もるのには十分です。このサービスは、25Gbや40Gbのリンクが1つだと提供できませんが、理論的には100 GbEリンクで十分です。ただし、ネットワーク冗長性が推奨されるため、このノードには少なくとも40Gbリンクが2つ必要で、標準的な4 x 25 GbE構成が推奨されます。

注意：コンポーネント ドライブの理論上のスループットのみに基づいてスループットを計算すると、1つのノードに対する見積もりが過度に高くなる可能性があります。**ノード上のRAIDコントローラーまたはHBAが、基盤となるストレージ メディアの最大スループットを満たしているか、または超えていることを確認します。**

7.2.2 書き込み負荷の高い環境

読み取り/書き込み操作で、PowerFlex環境にはさまざまなトラフィック パターンが生成されます。ホスト（SDC）で4k読み取り要求を1つ行う場合、1つのSDSにコンタクトしてデータを取得することになります。4kブロックは1つのSDSから1回転送されます。そのホストで4k書き込み要求を1つ行う場合、4kブロックはプライマリーSDSに転送され、続いてプライマリーSDSからセカンダリーSDSへコピーされることになります。

したがって、書き込み操作では、必要とするSDSの帯域幅が読み取り操作2倍になります。ただし、書き込み操作にはSDSが2つ関与しており、これに対し読み取り操作で必要なのは1つです。したがって、帯域幅要件の比率は、読み取り対書き込みで1:1.5です。

言い換えれば、SDSあたりで、書き込み処理に必要なネットワーク スループットは、基盤となるストレージのスループットと比較すると、読み取り操作の1.5倍になります。

通常の状況では、先に説明したストレージ帯域幅の計算で十分です。**ただし、環境内のSDSに、書き込み負荷の高いワークロードをホストすることが見込まれる場合は、ネットワーク容量を追加することを検討してください。**

7.2.3 ボリュームが別のシステムへレプリケートされている環境

バージョン3.5では、ネイティブの非同期レプリケーションが導入されています。このことは、帯域幅が生成されるのが、まずクラスター内で、次にレプリカ ピア システム間であることを考慮すると、説明がつきます。

7.2.3.1 レプリケートしているシステム内での帯域幅

先に述べたように、ボリュームがレプリケートされると、I/OがSDCからSDRへ送信され、その後に、ソース システムにはSDRからSDSへの後続I/Oが2つ発生します。SDRからはまず、関連するSDSへボリュームI/Oを移して処理（圧縮など）し、ディスクへ引き渡します。関連付けられているSDSは、SDRと同じノードにはない可能性があるため、このことを帯域幅の計算で考慮する必要があります。次の手順で、SDRでは受信する書き込みをジャーナリング ボリュームへ適用します。ジャーナル ボリュームはPowerFlexシステム内では別のボリュームのようになっているため、SDRではI/Oを、ジャーナル ボリュームが存在するストレージ プールをバックアップするさまざまなSDSへ送信しています。この手順では、追加的なI/Oが2つ追加されます。まずSDRで、ジャーナル ボリュームをバックアップする関連プライマリーSDSに対する書き込みをし、そのプライマリーSDSが、セカンダリーSDSへコピーを送信するためです。最後に、SDRでは、ジャーナル ボリュームから追加の読み取りを行ってから、リモート サイトへ送信します。

したがって、レプリケートされたボリュームの書き込み操作では、ソース クラスター内で必要になる帯域幅が、レプリケートされていないボリュームの書き込み操作の3倍になります。**レプリケートされたボリューム上で実行するワークロードの書き込みプロファイルは、慎重に検討します。追加の書き込みオーバーヘッドに対応するため、追加のネットワーク容量が必要になります。**したがって、システムをレプリケートする際は、バックエンドのストレージ トラフィックに対応するため、4 x 25 GbEまたは2 x 100 GbEのネットワークを使用することが推奨されます。

7.2.3.2 レプリカ ピア システム間での帯域幅

繰り返しますが、レプリカ ピア システム間でのネットワーク要件について検討するには、**ソース システムとターゲット システムとの間のレイテンシーが200msを超えないようにしてください。**

ジャーナル データは、ソースSDRとターゲットSDRとの間で送出されます。まず、レプリケーション ペア初期化フェーズで、次に、レプリケーション定常状態フェーズ時に行われます。ソースSDRとターゲットSDRとの間で十分な帯域幅を確保するためには、LANでもWANでも、特別に注意を払ってください。使用可能な帯域幅を超える可能性は、WAN接続よりも高くなります。書き込みフォールディングにより、ターゲット ジャーナルに送出するデータの量が低減することがありますが、これは必ずしも簡単に予測できるとは限りません。使用可能な帯域幅を超えた場合、ジャーナル インターバルはバックアップされ、ジャーナル ボリュームサイズとRPOのいずれも増大します。

ベスト プラクティスとして、レプリケートされるボリュームすべての持続的な書き込み帯域幅は、使用可能なWAN帯域幅の合計の80%を超えないようにすることが推奨されます。ピア システム同士がボリュームを相互にレプリケートしている場合、ピアSDR←→SDR帯域幅には、同時に双方向という要件を考慮する必要があります。特定のワークロードに必要なWAN帯域幅を計算するためのさらなるヘルプについては、最新のPowerFlex Sizerを参照して使用してください。

注意：Sizerツールは、Dellの従業員とパートナー様が利用できる社内ツールです。それ以外のユーザーは、WAN帯域幅のサイジングに関するサポートが必要な場合、自分のところのテクニカル セールス スペシャリストに相談する必要があります。

7.2.3.3 レプリケーション正常性についてのネットワーキングの含意

本書では、PowerFlexネットワーキング情報のベスト プラクティスに重点を置いていますが、ストレージ レイヤー自体の通常の動作、正常性、パフォーマンスは、導入されたネットワークの品質と容量によって異なります。これは特に、非同期レプリケーションやジャーナル ボリュームのサイジングとの関連性があります。

書き込みピークが推奨の「0.8 * WAN帯域幅」を超える可能性はありますが、短時間で済むようにしてください。ジャーナル サイズは、このような書き込みピークを吸収するのに十分になるよう大きくする必要があります。

このことは重要です。ジャーナル ボリューム容量は、ピア システム間のリンク アウテージに対応できるようサイジングする必要があります。アウテージは1時間と予測するのは合理的かもしれませんが、ユーザーには計画で3時間とすることを強くお勧めします。アウテージ中のアプリケーション書き込みを考慮して、十分なジャーナル スペースを確保する必要があります。**一般に、ジャーナル容量は、ピーク書き込み帯域幅*リンクのダウン タイムとして計算する必要があります。**最も繁忙な時間には、アプリケーション書き込み帯域幅を最大にする必要があります。ここでは、アプリケーションがピーク書き込みスループット1 GB/sに達しているとします。3時間とは10800秒です。したがって、必要なジャーナル容量は

$$1\text{GB/s} * 10800 \text{ seconds} = \sim 10.55\text{TB}$$

ただし、PowerFlexでは、ジャーナル容量をプール容量に対する割合で設定します。200 TBのストレージ プールが1つあると仮定します。

$$100 * 10.55\text{TB} / 200\text{TB} = 5.27\%$$

安全マージンとして、これを6%に丸めます。

注意：ジャーナル インターバルに送出されたボリューム データは圧縮されていません。PowerFlexでの圧縮は、静止データを対象としています。精細粒度ストレージ プールで、データ圧縮が行われるのはSDSサービスで、SDC（レプリケートされていないボリュームの場合）またはSDR（レプリケートされているボリュームの場合）から受信した後です。SDRは、レプリカ ペアのいずれの側のデータ レイアウトについても認識しておらず、それに依存していません。デスティネーション（ターゲット）のボリュームが圧縮済みに設定されている場合、ジャーナル インターバルが適用されると、ターゲット システムSDSで圧縮が行われます。

7.2.4 ハイパーコンバージド環境

PowerFlexがハイパーコンバージド導入環境にある場合、各物理ノードでは、SDS、ハイパーバイザー上のSDC、1つ以上のVMを実行しています。この意味で、ハイパーコンバージドPowerFlexの導入には、ハイパーバイザーが関与する必要はありません。ハイパーコンバージド導入環境では、ハードウェアへの投資が最適化されますが、ネットワークのサイジング要件も発生します。

先に説明したストレージ帯域幅の計算は、ハイパーコンバージド環境に適用されますが、仮想マシン、ハイパーバイザーまたはOSのトラフィック、SDCからのトラフィックに対するフロントエンド帯域幅も考慮する必要があります。仮想マシンのサイジングは、このテクニカル レポートの範囲外ですが、優先事項です。

ハイパーコンバージド環境では、ストレージを他のネットワーク トラフィックと論理的に分離することも優先事項です。

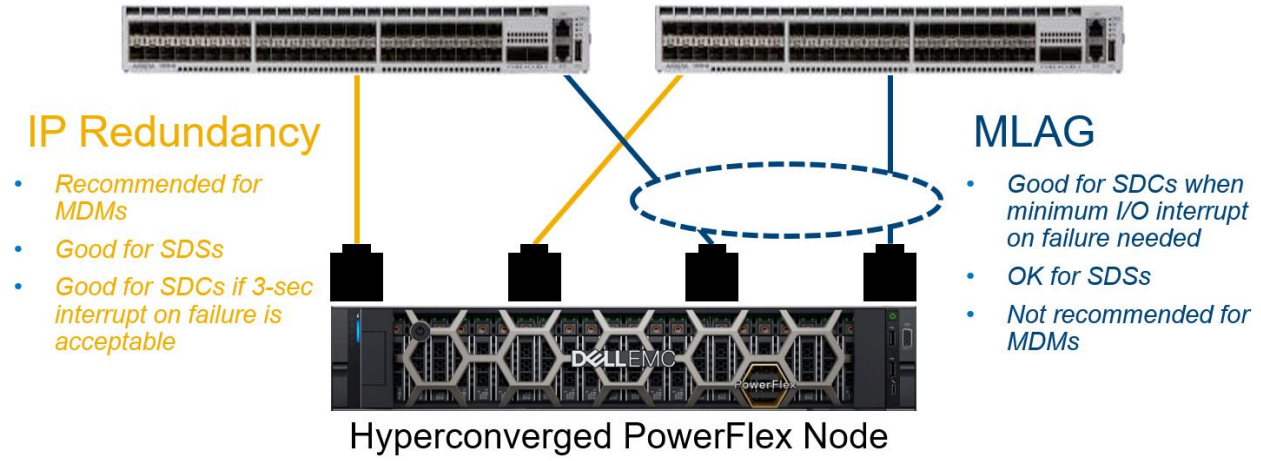

図 8　4 x 25Gbネットワーク接続を使用したハイパーコンバージドVMware環境の例を示します。このホストのPowerFlexトラフィックでは、ポートEth0とEth1を使用しています。冗長性は、MLAGではなく、ネイティブのPowerFlex IPマルチパス化によって実現されます。ポートEth2とEth3では、MLAGとVLANの両方のタグ付けを使用して、ハイパーバイザーなどのゲストにネットワーク アクセスを提供します。PowerFlexではVLANタグ付けとリンク アグリゲーションをサポートしているため、これ以外の構成も可能です。

8 ネットワーク ハードウェア

8.1 専用NIC

PowerFlexのエンジニアリングでは、可能であれば、PowerFlexトラフィックに専用ネットワーク アダプターを使用することが推奨されます。専用ネットワーク アダプターがあると、専用の帯域幅が得られ、トラブルシューティングがシンプル化します。共有ネットワーク アダプターがサポートされていること、ハイパーコンバージド環境では必須となることに注意してください。

8.2 共有NIC

最適ではありませんが、共有NICの使用はPowerFlexソフトウェアによってサポートされています。PowerFlexトラフィックが物理ネットワークを他の非PowerFlexトラフィックと共有する場合、QoSを実装して、PowerFlexトラフィックと非PowerFlexトラフィックのいずれかに起因するネットワーク混雑状態やスタベーションの問題を回避する必要があります。

8.3 2 x NIC対4 x NICとその他の構成

PowerFlexでは、追加的なネットワーク インターフェイスを追加することにより、ネットワーク リソースの拡張を行うことができます。**必須ではありませんが、ストレージ ネットワークでフロントエンドとバックエンドのトラフィックを独立させるのが理想的となる状況があります。**2層で導入することは、ストレージ チームや仮想化チーム、コンピューティング チームがそれぞれ独自のネットワークを管理する場合に役に立つことがあります。通常の場合、ユーザーはフロントエンドとバックエンドのネットワーク トラフィックをセグメント化して、ストレージ関連やアプリケーション関連のネットワーク トラフィックのパフォーマンスを保証します。いずれの場合も、冗長性、容量、速度のためには、インターフェイスを複数使用することが推奨されます。

PCI NICの冗長性も考慮事項です。**デュアルポートPCI NICを各サーバーで2つ使用する方が、クアッドポートPCI NICを1つ使用するよりも好ましいです。デュアルポートPCI NICを2つ構成しておくと、1つのNICでの障害を乗り切ることができるためです。**

8.4 スイッチの冗長性

ほとんどのリーフ/スパイン構成では、スパイン スイッチとトップオブラック（ToR）リーフ スイッチが冗長化されています。これにより、ToRスイッチに障害が発生した場合でも、ネットワークのラック内部のコンポーネントへのアクセスは継続します。各ラックに搭載されているToRスイッチが1つの場合、ToRスイッチに障害が発生すると、ラック内部のSDSコンポーネントにアクセスできなくなります。**したがって、ToRスイッチが1つという構成は推奨されません。**ToRスイッチがラックごとに1つ使用される場合、ユーザーはフォールト セットをラック レベルで定義して、スイッチに障害が発生した場合のデータの可用性を確保する必要があります。

9 IPに関する考慮事項

9.1 IPv4とIPv6

バージョン2.6以降、3.0より後はすべてのバージョンに含まれていますが、PowerFlexでは、2層とハイパーコンバージドの両方の導入オプションで、IPv6サポートを提供します。以前のバージョンのPowerFlexでは、Internet Protocol version 4（IPv4）のアドレス指定のみがサポートされていました。本書の例では、IPv4に重点を置いています。

9.2 IPレベルの冗長性

MDM、SDS、SDR、SDCはIPアドレスを複数持つことができるため、複数のネットワークに存在することが可能です。これにより、ロード バランシングと冗長性のオプションが得られます。

PowerFlexでは、ソフトウェア コンポーネントが複数のリンクでトラフィックを送信するよう設定されている場合、物理ネットワーク リンク全体にわたる冗長性とロード バランシングをネイティブで提供します。この設定では、MDM、SDR、またはSDSで利用可能な各物理ネットワーク ポートに、独自のIPアドレスが、それぞれ別のサブネットで割り当てられます。

サブネットを複数使用すると、冗長性がネットワーク レベルで得られます。また、サブネットを複数使用することで、トラフィックが1つのコンポーネントから別のコンポーネントへ送信されるようになります。これにより、ソース コンポーネントのルート テーブルでは、デスティネーションIPアドレスに応じて、異なったエントリーが選択されます。これにより、ソースの1つの物理ネットワーク ポートがボトルネックになることがなくなります。ソースがコンタクトする複数のIPアドレス（それぞれ物理ネットワーク ポートに対応）が1つのデスティネーションにあるためです。

言い換えれば、ソースとデスティネーションの複数の物理ポートが同じサブネット内にある場合、ソース ポートでボトルネックが発生する場合があります。たとえば、2つのSDSが1つのサブネットを共有している場合、各SDSには2つの物理ポートがあり、各物理ポートには独自のIPアドレスがそのサブネットにあり、IPスタックによってソースSDSは常に同じ物理ソース ポートを選択します。**サブネット全体でポートを分割すると、各ポートはホストのルーティング テーブルにあるそれぞれ別のサブネットに対応するため、ロード バランシングが可能になります。**

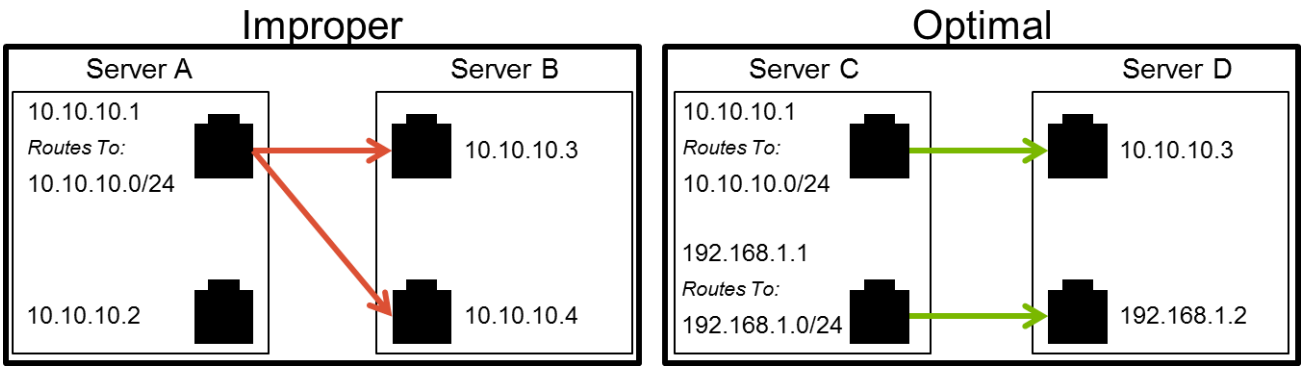

図 9　オペレーティング システムのIP設定の比較。左側の不適切なIP設定では、すべてのトラフィックに対して同じサブネット（10.10.10.0/24）が使用されます。サーバーAがサーバーBへの接続を開始すると、送信接続には常に、10.10.10.0/24へのルートを提供するネットワーク リンクが選択されます。サーバーAの2つ目のネットワークポートは、送信接続には使用されません。右側の適切なIP設定では、2つのサブネット、10.10.10.0/24と192.168.1.0/24が使用されます。これにより、サーバーCのいずれのポートも送信接続に使用することができます。注意：この例で選択されているサブネット（10.10.10.0/24と192.168.1.0/24）は任意です。クラス「A」とクラス「C」を同時に使用するのは、視覚的に区別するためです。

MDMまたはSDSのそれぞれが複数のIPアドレスにアクセスできる場合、PowerFlexではトラフィック パターンを認識しているため、ロード バランシングをより効果的に処理します。これにより、パフォーマンスが少し向上する可能性があります。さらに、リンク アグリゲーションでは、リンクレベルのフェールオーバー用に、独自のタイマー セットを保持します。したがって、ネイティブPowerFlexにIPレベルの冗長性があると、リンクがダウンしたときのトラブルシューティングが容易になります。

また、IPレベルの冗長性によって、IPアドレスの競合からも保護します。IPの不要な変更や競合を防止するため、**PowerFlexのMDMまたはSDCが存在するネットワークには、DHCPを導入しないでください。**

独立して使用する場合、MDM to MDM通信に使用するリンクには、MLAGよりもIPレベルの冗長性が強く推奨されます。IPレベルの冗長性が、VLANの冗長リンク アグリゲーション グループで一番上のレイヤーになっている場合、両方のテクノロジーを使用するのが適切です。この例については、『Dell EMC PowerFlex Appliance Network Planning Guide』を参照してください。

10 Ethernetに関する考慮事項

10.1 ジャンボ フレーム

PowerFlexではジャンボ フレームをサポートしており、ストレージ トラフィックにはジャンボ フレームを使用することが非常に推奨されます。ただし、ジャンボ フレームを有効化すると、ネットワーク インフラストラクチャによっては困難が生じることがあります。さまざまなネットワーク コンポーネントによるジャンボ フレームの実装に一貫性がないと、トラブルシューティングが困難になるというパフォーマンスの問題が発生することがあります。ジャンボ フレームを使用するには、PowerFlexインフラストラクチャ（ホストとスイッチを含む）と（HCIが導入されている場合は）ストレージVMで使用されているすべてのネットワーク コンポーネントで有効化する必要があります。

ジャンボ フレームを有効化すると、1つのEthernetフレームで、より多くのデータを渡すことができます。これにより、Ethernetフレームの総数と、各ノードで処理する必要がある割り込みの数が減少します。PowerFlexインフラストラクチャのすべてのコンポーネントでジャンボ フレームが有効化されていると、ワークロードによっては、パフォーマンス上のメリットが約10%向上します。

注意：PowerFlex Managerを使用してPowerFlexクラスターをアプライアンスまたはラック システムに導入する場合、ノードとスイッチ コンポーネントでのジャンボ フレームの設定は、すべてのクラスター コンポーネントに合わせて全体が調整され管理されます。

ネットワーク コンポーネントを注意深く確認し、ジャンボ フレームの設定が各ポイントで一貫しているようにします。不確実な場合は、ジャンボ フレームを初めは無効化しておくことが推奨されます。ジャンボ フレームを有効化するのは、動作のセットアップが安定して、その使用がインフラストラクチャでサポートできることを確認してからにしてください。ジャンボ フレームがすべてのノードで各パスに沿って設定されているようにするには、Linuxの`tracepath`コマンドなどのユーティリティーを使用して、MTUサイズをパスに沿って探索します。pingは、ジャンボ フレームの問題の診断にも有用です。Linuxで使用するコマンドの形式は、`ping -M do -s 8972 <ip address/hostname>`です。（ここでは、MTUサイズ9000から、カプセル化されていないパケット ヘッダー28バイトを減算していることに注意してください。）

ジャンボ フレームの実装の詳細については、『PowerFlex Configure and Customize guide』を参照してください。

10.2 VLANタグ付け

PowerFlexは、サーバーとアクセスまたはリーフ スイッチとの間の接続で、ネイティブVLANとVLANタグ付けに依存していません。これらは、オペレーティング システムまたはスイッチで設定されている場合、PowerFlexソフトウェアに対して透過的です。VLANは、PowerFlexエンジニアリングによって測定すると、パフォーマンス レベルに影響を及ぼしません。

当社はPowerFlexアプライアンスの導入で、一連のVLANが均一に設定されていることを想定しています。後述のセクション19を参照してください。

11 リンク アグリゲーション グループ

リンク アグリゲーション グループ（LAG）とマルチシャーシ リンク アグリゲーション グループ（MLAG）では、ポートをエンドポイント間で結合します。エンドポイントは、スイッチ、LAGまたは2つのスイッチを搭載したホスト、MLAGを搭載したホストにすることができます。リンク アグリゲーションの用語と実装方法は、スイッチ ベンダーによって異なります。Cisco Nexusスイッチでは、MLAG機能は仮想ポート チャネル（vPC）と呼ばれています。

LAGでは、リンク アグリゲーション制御プロトコル（LACP）を使用して、セットアップ、ティアダウン、エラー処理を行います。LACPは標準ですが、独自仕様のバリアントが多数あります。

スイッチ ベンダーや、PowerFlexをホストするオペレーティング システムに関係なく、**リンク アグリゲーション グループを使用する場合はLACPが推奨されます。スタティック リンク アグリゲーションの使用はサポートされていません。**

リンク アグリゲーションは、各物理ポートに独自のIPアドレスが割り当てられていれば、IPレベルの冗長性の代替として使用できます。リンク アグリゲーションは、一部のチームにとっては設定しやすくなっており、IPアドレスの枯渇が問題になっている場合にも有用です。リンク アグリゲーションは、PowerFlexが実行されているノードと、アタッチされているネットワーク機器との、両方に設定する必要があります。

PowerFlexは、IPレベルの冗長性とリンク アグリゲーションのいずれを選択するかにかかわらず、耐障害性とハイ パフォーマンスを実現します。MLAGが使用されている場合、SDSのパフォーマンスはIPレベルの冗長性のパフォーマンスに近くなります。

- **SDSで、MLAGかIPレベルの冗長性かを選択するにあたっては、運用上の意思決定を考慮する必要があります。**
- **MDM to MDMトラフィックでは、MLAGよりもIPレベルの冗長性またはLAGが強く推奨されます。MDM上の1つのIPアドレスの可用性が継続すると、フェールオーバーが回避されるためです。MDMは複数のIPアドレス間で通信を行うよう設計されているので、MDM間のタイムアウトは短いのです。**
- **3.5ではネットワークの耐障害性が向上したため、SDCコンポーネントで使用されるリンクでは通常、MLAGよりもIPレベルの冗長性が推奨されます。**

11.1 LACP

LACPでは、ネットワーク リンクのアグリゲートされたグループにある各物理ネットワーク リンクにわたって、メッセージを定期的に送信します。このメッセージは、各物理リンクがまだアクティブかどうかを判断するロジックの一部です。これらのメッセージの頻度は、ネットワーク管理者がLACPタイマーを使用して制御することができます。

LACPタイマーは通常、リンク障害を高速で（1秒あたり1メッセージを）、または通常の速度で（30秒ごとに1メッセージを）検出するよう設定できます。LACPタイマーが高速で動作するよう設定されている場合は、対応処置も迅速に取られます。さらに、メッセージを毎秒送信する場合の相対的なオーバーヘッドは、最新のネットワーク テクノロジーでは小さくなっています。

LACPタイマーは、リンク アグリゲーションがPowerFlex SDSとスイッチとの間で使用される場合、高速で動作するよう設定する必要があります。

LACP接続を確立するには、LACPピアの一方または両方でアクティブ モードを使用するよう設定する必要があります。**したがって、PowerFlexノードに接続されているスイッチには、アクティブ モードをリンク全体で使用するよう設定することが推奨されます。**

11.2 ロード バランシング

リンク アグリゲーション グループで複数のネットワーク リンクがアクティブになっている場合、エンドポイントでは、リンク間でトラフィックを分散する方法を選択する必要があります。ネットワーク管理者は、エンドポイントにロード バランシング方法を設定することで、この動作を制御します。ロード バランシング方法では通常、ソースまたはデスティネーションのIPアドレス、MACアドレス、またはTCP/UDPポートの何らかの組み合わせに基づいて、使用するネットワーク リンクを選択します。

このロード バランシング方法は、「ハッシュ モード」といいます。ハッシュ モードのロード バランシングは、同じ物理リンク上のソース アドレスとデスティネーション アドレスまたはトランスポート ポートの特定のペアとの間で （リンクがアクティブである限り） トラフィックを保持することを目的としています。

ハッシュ モードのロード バランシングで推奨される設定は、使用しているオペレーティング システムによって異なります。

SDSを実行しているノードがスイッチへのリンクをアグリゲートしておりVMware ESX®を実行している場合、ハッシュ モードは、「ソースIPアドレスとデスティネーションIPアドレス」または「ソースIPアドレスとデスティネーションIPアドレスとTCP/UDPポート」を使用するよう設定する必要があります。

SDSを実行しているノードがスイッチへのリンクをアグリゲートしておりLinuxを実行している場合、Linux上のハッシュモードは、ボンディング オプション「xmit_hash_policy=layer2+3」または「xmit_hash_policy=layer3+4」を使用するよう設定する必要があります。ボンディング オプション「xmit_hash_policy=layer2+3」では、ソースとデスティネーションのMACアドレスとIPアドレスを使用してロード バランシングを行います。ボンディング オプション「xmit_hash_policy=layer3+4」では、ソースとデスティネーションのIPアドレスとTCP/UDPポートを使用してロード バランシングを行います。

Linuxでは、「miimon=100」ボンディング オプションも使用する必要があります。このオプションを選択すると、Linuxでは各物理リンクのステータスを100ミリ秒ごとに検証します。

各ボンディング オプションの名前はLinuxディストリビューションによって異なる場合がありますが、推奨事項はいずれも同じです。

11.3 マルチシャーシ リンク アグリゲーション グループ

リンク アグリゲーション グループ（LAG）と同様、MLAGではネットワーク リンクの冗長性を実現します。LAGと異なり、MLAGでは、1つのエンドポイント（PowerFlexを実行しているノードなど）を複数のスイッチへ接続することができます。MLAGを参照する際に使用する名前は、スイッチ ベンダーによって異なり、MLAGの実装は通常、独自仕様となっています。

MLAGの使用はPowerFlexでサポートされていますが、MDM to MDMトラフィックには通常推奨されません。ただし、次のセクションにある注意を参照してください。「ロード バランシング」セクションに記載されているオプションは、MLAGの使用にも適用されます。

12 MDMネットワーク

MDMは、ホスト（SDC）とその分散ストレージ（SDS）との間のデータ パスに存在するわけではありませんが、それらの間での関係を保持し、クラスターの状態を常時追跡する役割を担います。したがって、MDM to MDMトラフィックは、MLAGでの物理ネットワーク リンク ロスなど、レイテンシーに影響を与えるネットワーク イベントに敏感です。

MDMは冗長化されています。したがってPowerFlexは、レイテンシーの増大だけでなく、MDMロスがあってもサバイバルできます。MDMをホストしているノードにも、MLAGの使用は奏功します。**ただし、MDM to MDMトラフィックを実行するネットワーク上でMLAGを使用する必要がある場合は、Dell EMC PowerFlexの担当者と協力して、ネットワーク冗長性を2倍にした堅牢な設計を選択して、MLAGとネイティブのIPレベルの冗長性とを組み合わせるようにしてください。**

ほとんどの状況で、MDMでは、MLAGでなく、2つ以上のネットワーク セグメントでIPレベルの冗長性を使用することが推奨されます。MDMでは、1つ以上の専用MDMクラスター ネットワークを共有する可能性があります。

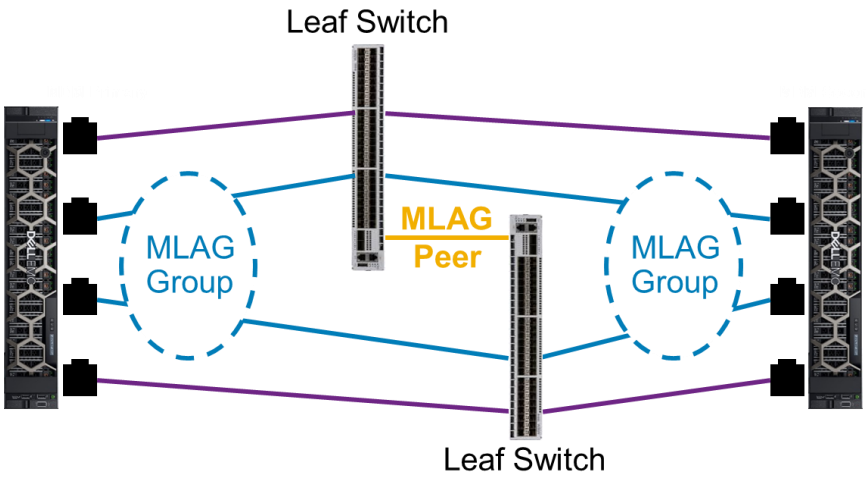

図 10　ノード2つがリーフ スイッチ2つに接続されています。MDMトラフィックは、紫色のリンクをトラバースする必要があります。当該リンクはMLAGグループに属していないためです。

13 ネットワーク サービス

13.1 DNS

MDMクラスターでは、システム コンポーネントとそのIPアドレスのデータベースのメンテナンスをします。PowerFlex導入環境に影響を与えるDNSアウテージの可能性を排除するため、MDMクラスターでは、ホスト名または完全修飾ドメイン名（FQDN）によるシステム コンポーネントの追跡はしません。MDMクラスターでシステム コンポーネントを登録するときにホスト名またはFQDNを使用すると、IPアドレスに解決され、コンポーネントにはそのIPアドレスが登録されます。

これについての例外は、VASAプロバイダーが導入され、vVolが実装されている場合です。PowerFlex環境でvVolを使用するには、PowerFlex VASAプロバイダー（シングル モードと3ノード クラスターのいずれか）を導入する必要があります。vVolテクノロジーをvSphere環境へ導入するには、vCenter Server、vVolデータストアを使用するESXiホスト、VASAプロバイダー ホストそれ自体の完全なFQDNが必要です。これらのコンポーネントすべてに有効なDNS解決がなければなりません。したがって、高可用性のDNSサービスを採用して、vVolの接続と機能のロスを回避する必要があります。

要約すると、**ホスト名やFQDNの変更は通常の場合、vVolが実装されていない限り、PowerFlex導入環境でのコンポーネント間のトラフィックに影響を及ぼしません。**

14 WAN上でのレプリケーション ネットワーク

PowerFlexネイティブ非同期レプリケーションを使用する場合に考慮すべき追加の考慮事項があります。セクション2.4と3.9では、Storage Data Replicator（SDR）とそのトラフィックを取り上げました。セクション7.2.3では、帯域幅の追加的要件を取り上げました。このセクションでは、ワイド エリア ネットワーク（WAN）上で実行されるレプリケーションに特化して、アドレス指定とルーティングのトピックを検討します。推奨事項は一般論であり、実装の詳細はハードウェアとWANで使用されるトポロジーによって異なります。

14.1 追加的IPアドレス

保護ドメイン内では、SDRはSDSと同じホストにインストールされますが、ジャーナル ボリュームに対してSDRによる書き込みが行われるトラフィックは、ジャーナルをホストしているすべてのSDSへ送信されます（ジャーナルはホスト上に共存するものだけではありません）。バックエンドのストレージ ネットワークの各SDRは、SDSと同じノードのIPでリッスンするため、保護ドメインのすべてのSDSにアクセスできることが必要です。

ただしSDRでは、リモートSDRとの通信を可能にする追加的IPアドレスが別途必要です。ほとんどの場合、適切に設定されたゲートウェイによりルーティングできるアドレスにする必要があります。冗長性を確保するため、SDRごとに2つ必要です。

14.2 ファイアウォールに関する考慮事項

SDRは相互に通信し、その間でレプリケートされたデータをTCPポート1088経由で送出します。このポートは、ソース システム側のファイアウォールの出口に向けて開いている必要があります。また、ターゲット システム側の入口に向けて開いている必要があります。レプリケーションが2つのシステムの間で双方向に実行されている場合、ポート1088はファイアウォールで両側の出口と入力の両方に向けて開いている必要があります。

14.3 スタティック ルート

PowerFlex非同期レプリケーションは通常、WAN上で同じアドレス セグメントを共有せず物理的にリモートのクラスター間で発生します。デフォルト ルートそのままではリモートのSDR IPへ適切にパケットを移動させるのに適していない場合、スタティック ルートは、次のホップ アドレスまたは出口インターフェイスのいずれかあるいは両方を示してリモート サブネットに到達する設定する必要があります。

例：`X.X.X.X/X via X.X.X.X dev interface`

いずれの側にも数個のノードがある小規模システムを検討してください。各ノードにはネットワーク アダプターが4つ搭載されており、そのうち2つはPowerFlexクラスター内部での通信用のIPが設定されており、もう2つは、サイト間の外部通信用のIPアドレスが設定されています。

この例では、ノードに対し、指定のゲートウェイを介して、もう一方の側のWANサブネットにアクセスするよう指示しています。ソース サイトAからは、ネットワーク インターフェイスenp130s0f0およびenp130s0f1に、それぞれ範囲が30.30.214.0/24および32.32.214.0/24のアドレスが設定されます。それぞれのルート インターフェイスファイルを設定すると、指定のゲートウェイとインターフェイスを介して、リモート ネットワークのパケットを移動させることができます。

```
route-enp130s0f0コンテンツ→        31.31.0.0/16 via 30.30.214.252 dev enp130s0f0

route-enp130s0f1コンテンツ→        33.33.0.0/16 via 32.32.214.252 dev enp130s0f1
```

リモート ネットワーク31.31.214.0/24に向けられたパケットは、ゲートウェイIP 30.30.214.252の次のホップ アドレスを経由して移動します。デスティネーションが33.33.214.0/24のパケットについても同様です。

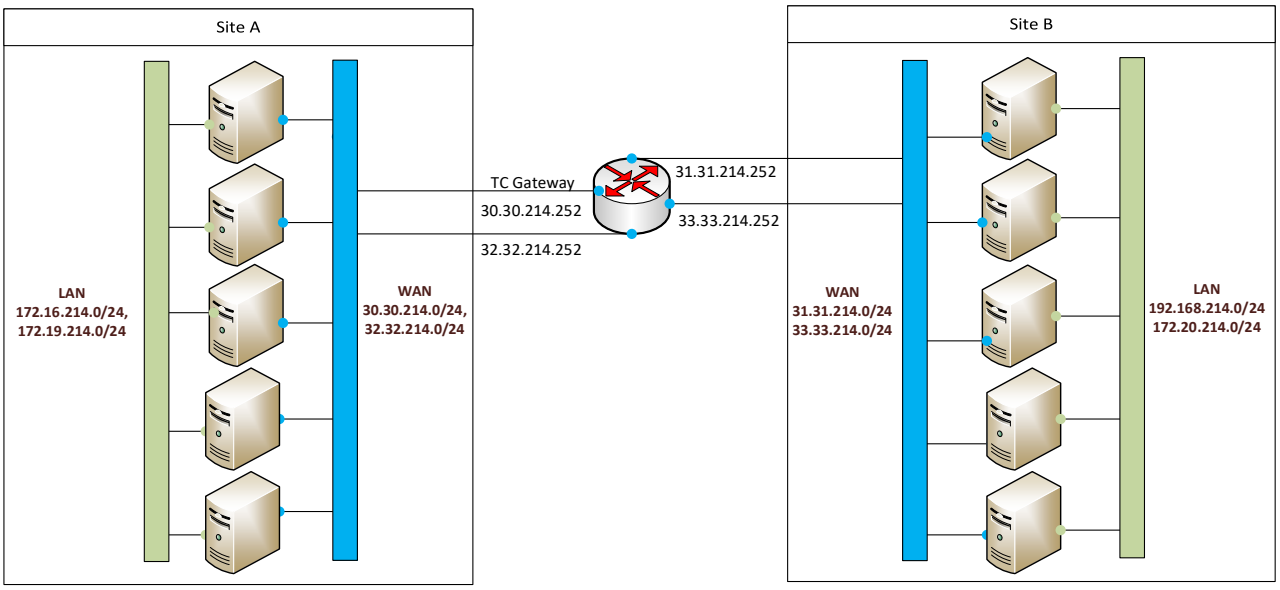

図 11　　PowerFlexレプリケーションのWANトポロジーの例。

スタティック ルート設定の詳細は、ご使用のオペレーティング システム/ハイパーバイザーと全体的なネットワーク アーキテクチャによって異なりますが、一般原則は同じです。

14.4　MTUとジャンボ フレーム

MTUは、SDR間のネットワーク インターフェイスで、WANリンク設定と一致させるよう、適正に設定する必要があります。多くの場合、これは1500となります。これは、ジャンボ フレームがすべてのローカル ネットワークでパフォーマンス強化として有効化されている場合、特に重要で覚えておくべきことです。MTUがWAN設定と一致しない場合にIPの断片化が発生すると、レプリケーション パフォーマンスが低下します。ハードウェア構成によっては、MTUの不整合により、インターフェイスに到達したときにパケットが全部ドロップされることがあります。したがって、どのような場合でも、WANのMTUは既知、かつテストされている必要があります。

15 動的ルーティングに関する考慮事項

数百のノードで構成される大規模なリーフ/スパイン環境では、PowerFlexトラフィックを動的ルーティングするネットワーク インフラストラクチャが必要になることがあります。

PowerFlexトラフィックをルーティングする中心的な目的は、ルーティング プロトコルのコンバージェンス時間を短縮することです。コンポーネントまたはリンクに障害が発生した場合、ルーターまたはスイッチで障害を検出する必要があります。ルーティング プロトコルでは、変更を他のルーターへ伝播する必要があります。そして、各ルーターまたはスイッチでは、各デスティネーション ノードへのルートを再計算する必要があります。ネットワークが正しく設定されている場合、このプロセスは300ミリ秒未満で発生します。これは、MDMクラスターの安定性を維持するのに十分な速度です。

極端な混雑状態やネットワーク障害の発生時に、コンバージェンス時間が400ミリ秒を超えた場合、MDMクラスターはセカンダリーMDMへフェールオーバーされることがあります。MDMがフェールオーバーした場合でも、システムの動作は続行し、**システムの最大限の安定性を維持するための目標が300ミリ秒**であっても、I/Oは続行します。他のシステム コンポーネント通信メカニズムのタイムアウト値ははるかに高くなっているため、システムの設計で、最も要求の厳しいタイムアウト要件（MDMのそれ）に合わせる必要があります。

コンバージェンス時間を可能な限り最短にするにはは、標準的なベスト プラクティスが適用されます。これは、その目的（パワー不足のルーター（脆弱なリンク）がなければ急速なコンバージェンスを防ぐことができるなど）を達成するために設計されたすべてのネットワーク ベンダーのベスト プラクティスに準拠することを意味します。

テスト済みのどのネットワーク ベンダーも、デフォルトのOSPFまたはBGP設定だと、コンバージェンス時間が不十分です。**ネットワーク ベンダーに関係なく、ルーティング プロトコル導入のたびに、コンバージェンス時間を最短にするためのパフォーマンス調整を含める必要があります。**この調整には、双方向転送検出（BFD）の使用と、障害に関連するタイミング メカニズムの調整が含まれます。

OSPFとBGPの両方とも、PowerFlexでテスト済みです。PowerFlexは、ルーティング プロトコルとネットワーキング デバイスが適切に設定されていれば、リンクやデバイスでの障害発生時にもエラーなしで機能することが知られています。しかしながら、**BGPよりもOSPFが推奨されます。**この推奨事項は、OSPFとBGPの両者が急速なコンバージェンスに向けて最適に設定されている場合に、OSPFはBGPよりもコンバージェンスが急速であることを示すテスト結果によって支持されています。

15.1 BFD（双方向転送検出）

ルーティング プロトコル（OSPFまたはBGP）のいずれを選択するかにかかわらず、BFD（双方向転送検出）を使用する必要があります。BFDによって、プロトコルネイティブのhelloタイマーに関連するオーバーヘッドが削減され、リンク障害が迅速に検出されるようになります。BFDは、ルーターによるCPUや帯域幅の使用率を低減するなど多くの理由により、ネイティブのプロトコルhelloタイマーよりも高速な障害検出を実現します。**したがって、BFDは、アグレッシブなプロトコルhelloタイマーよりも強く推奨されます。**

PowerFlexは、BFDを搭載しOSPFやBGPのルーティングを最適化して導入してあれば、ネットワーク フェールオーバー時も安定しています。BFDを使用して、サブ秒障害検出を有効化する必要があります。

ネットワークのコンバージェンスでは、イベントが検出され、他のルーターへ伝播され、ルーターによって処理され、またルーティング情報ベース（RIB）または転送情報ベース（FIB）が更新される必要があります。ルーティング プロトコルのコンバージェンスでは、これらすべての手順が実行される必要があり、すべて300ミリ秒未満で完了することになっています。

Cisco 9000シリーズ スイッチを使用したテストでは、**150ミリ秒のBFDホールド ダウン タイマー**で十分でした。150ミリ秒のホールド ダウン タイマーに設定した内容は、伝送インターバルが50ミリ秒、min_rxが50ミリ秒、乗数は3です。PowerFlexでの推奨事項は、最大で150ミリ秒のホールド ダウン タイマーを使用することです。スイッチ ベンダーがサポートしているBFDホールドダウン タイマーが150ミリ秒未満の場合は、達成可能な中で最短のホールド ダウン タイマーが推奨されます。可能な場合は、BFDを非同期モードで有効化してください。

Cisco vPC（MLAG）を使用している環境では、**ルーティングされているすべてのインターフェイスと、ファースト ホップ冗長プロトコル（FHRP）を実行しているすべてのホストフェイシング インターフェイスで、BFDも有効化する必要があります。**

```
feature bfd

hsrp bfd all-interfaces

interface Vlan<num>
no shutdown
no ip redirects
ip address 192.168.103.2/24
no ipv6 redirects
hsrp version 2
hsrp 103
authentication text Vce12345
preempt
priority 110
ip 192.168.103.1

router ospf 1
  bfd
  bfd all-interfaces strict-mode

interface eth <x/x> / vlan <num> / Po <num>
 bfd interval 50 min_rx 50 multiplier 3
```

図 12　アグリゲーション–アクセス/スパイン リーフ トポロジーを使用したCiscoスイッチのBFD設定の一例。BFDには、150ミリ秒（インターバルは50マイクロ秒、乗数は3）のホールド ダウン タイマーが設定されています。インターフェイス ポート - channel51のOSPFと、インターフェイスVlan30のHSRPは、いずれもBFDのクライアントとして設定されています。

```
bfd ipv4 interval 50 min_rx 50 multiplier 3


interface Vlan30
  bfd interval 50 min_rx 50 multiplier 3
  no bfd echo
  vrrp 1
    vrrp bfd 30.30.30.124

interface port-channel49
  no bfd echo
  bfd per-link

interface port-channel51
  no bfd echo
  bfd per-link
router ospf 100
  bfd
```

図 13　アグリゲーション–アクセス トポロジーでのDell BFD設定の一例。BFDには、150ミリ秒（インターバルは50マイクロ秒、乗数は3）のホールド ダウン タイマーが設定されています。インターフェイス ポート - channel51のOSPFと、インターフェイスVlan30のVRRPは、いずれもBFDのクライアントとして設定されています。

これらの設定については、次の点に注意してください。

- ポート-チャネル インターフェイスの場合は、リンク単位のBFDを有効化する必要があります。
- BFDでは、IPリダイレクトを無効化する必要があります。（BFDの動作を確保するためのオーバーライド）
- FHRPは、アクセス-アグリゲーション トポロジーにのみ必要です。

15.2　物理リンクの構成

リンク障害に関連するタイマーは、チューニングの候補となります。リンク ダウンとインターフェイス ダウン イベントの検出と処理は、ネットワーク ベンダーと製品ラインによって異なります。**Cisco Nexusスイッチでは、各SVIインターフェイスの「`carrier-delay`」タイマーを100ミリ秒に設定する必要があります。また、各物理インターフェイスの「`link debounce`」タイマーを500ミリ秒に設定する必要があります。**

キャリア遅延（`carrier-delay`）は、スイッチのタイマーです。SVIインターフェイスに適用されます。キャリア遅延とは、リンク障害が検出されたときにスイッチがアプリケーションに通知する前に待機する時間量です。キャリア遅延は、不安定なネットワークでのフラッピング イベント通知を防止するために使用されます。最新のリーフ/スパイン環境では、すべてのリンクをポイントツーポイントとして設定し、ネットワークを安定させる必要があります。PowerFlexトラフィックを送信するSVIインターフェイスの推奨値は、100ミリ秒です。

デバウンス（`link debounce`）は、ファームウェアでリンクダウン通知を遅延させるタイマーです。物理インターフェイスに適用されます。デバウンスはキャリア遅延に似ていますが、論理インターフェイスではなく物理インターフェイスに適用され、リンクダウン通知のみに使用されます。待機期間中はトラフィックが停止します。link debounce設定を0以外にすると、ルーティング プロトコルのコンバージェンスに影響する可能性があります。リンク デバウンス タイマーの推奨値は、PowerFlexトラフィックを送信する物理インターフェイスの場合、500ミリ秒です。

```
interface vlan <num>
 carrier-delay msec 100

interface eth <x/x>
  link debounce time 500
```

15.3 ECMP

等コスト マルチパス ルーティング（.ECMP）を使用する必要があります。.ECMPは、トラフィックをリーフ/スパイン スイッチ間で均等に分散させ、冗長リーフからスパイン ネットワークへのリンクを使用して高可用性を実現します。ECMPはMLAGに類似していますが、Ethernet上ではなくレイヤー3（IP）上で動作します。

デフォルトでは、.ECMPはCisco NexusスイッチのOSPFでオンになっています。Cisco NexusスイッチのBGPのデフォルトでオンになっていない場合は、手動で有効化する必要があります。使用するECMPハッシュ アルゴリズムは、レイヤー3（IP）、またはレイヤー3とレイヤー4（IPとTCP/UDPポート）である必要があります。

15.4 OSPF

OSPFは、優先されるルーティング プロトコルです。適切に設定されている場合、コンバージェンスが急速になるためです。OSPFを使用する場合、リーフ/スパイン スイッチはすべて1つのOSPF領域にあります。**イントラMDM通信を安定させるには、コンバージェンス時間300ミリ秒未満が要求されます。**すべてのリーフ/スパイン スイッチで、OSPFインターフェイスは、BFDのクライアントとして設定されているOSPFプロセスで、ポイントツーポイントに設定する必要があります。これにより、タイマーが正しく設定されるようになります。デフォルトから変わりません。**さらに、ToR-Agg（Access-Agg）トポロジーでのL3ハンドオフについては、OSPFインターフェイスをポイントツーポイントに設定する必要があります。**

15.5 BGP

OSPFはコンバージェンスがより急速なので優先されますが、BGPも要求どおりのタイム フレーム内でコンバージェンスをするよう設定します。

デフォルトでは、BGPはCisco NexusスイッチでECMPを使用するようには設定されていません。手動で設定する必要があります。IBGPとEBGPのいずれも、デフォルトではECMPをサポートしていないため、設定する必要があります。IBGPの設定では、BGPルート リフレクターと追加パス機能が必要です。スパイン/リーフ トポロジーでECMPを完全にサポートするためです。

BGPは、それぞれのリーフ/スパイン スイッチが異なったASN（自律型システム番号）を表すように設定することができます。この設定での各リーフは、他のスパインそれぞれとピアになる必要があります。

リーフ/スパイン スイッチでは、ECMPも有効化して、スイッチが複数のBGPパスにわたってロード バランシングを行えるようにする必要があります。このことは、Ciscoでは、「`maximum-path`」パラメーターをスパイン スイッチへの使用可能なパス数に設定することも含まれています。

PowerFlexのBGPでは、BFDを各リーフ/スパイン ネイバーに設定する必要があります。BGPを使用している場合、**SDSネットワークとMDMネットワークはリーフ スイッチによってアドバタイズされます。**

リーフ設定

```
router bgp 100
  router-id 1.1.1.2
  address-family ipv4 unicast
    maximum-paths ibgp 3
    address-family l2vpn evpn
    maximum-paths ibgp 3

  neighbor 11.11.11.11
    bfd
    remote-as 100
    update-source loopback0
    address-family ipv4 unicast
      send-community
      send-community extended
    address-family l2vpn evpn
      send-community
      send-community extended

  vrf VxFLEX_MGMTanagement_VRF
    address-family ipv4 unicast
      maximum-paths ibgp 3
      advertise l2vpn evpn
      redistribute direct route-map ALL
```

スパイン設定

```
router bgp 100
  router-id 11.11.11.11
  address-family ipv4 unicast
   maximum-paths ibgp 3
  address-family l2vpn evpn
   maximum-paths ibgp 3

  neighbor 1.1.1.1
    bfd
    remote-as 100
    update-source loopback0
    address-family ipv4 unicast
    address-family l2vpn evpn
      send-community
      send-community extended
      route-reflector-client
```

図 14　Cisco Nexusリーフスイッチ（左）とスパイン スイッチ（右）のBGP設定例。これらは同じ自律型システム（100）にあります。「`maximum-path`」パラメーターは、ECMPに使用するパスの数と一致するようチューニングされます。（この例では3ですが、必ずしもそうではありません）。BFDは、リーフまたはスパインの各ネイバーで有効化されます。リーフ スイッチは、PowerFlexのMDMネットワークとSDSネットワークをアドバタイズするよう設定されています。

注意：

- スパイン/リーフ トポロジーを使用するPowerFlexラック システムでは、制御プレーンの通信とEVPNの到達可能性に向けて、BGPを使用します。データプレーンにはOSPFが使用されています。
- maximum-pathにより、複数のNVEインターフェイスでVTEPの到達可能性を実現
- IBGPには、スパインをルートリフレクターとして使用するよう設定
- BGP as-path multipath-relaxは、EBGPを使用していないため、適用されません。

15.6 リーフ/スパインの帯域幅要件

ストレージ メディアがパフォーマンスのボトルネックではないと仮定すると、リーフ/スパイン スイッチ間で必要となる帯域幅の量を計算するには、各リーフ スイッチからアタッチされているホストまでで使用可能な帯域幅の量を決定し、リーフ スイッチにローカルになるとみられるI/Oの量を割り引き、各スパイン スイッチ間でのリモート帯域幅要件を除算します。

ラックは2台で、各ラックにはリーフ スイッチ2個とサーバー20台があり、各サーバーには25Gbインターフェイスが2つあり、これらのサーバーはそれぞれラック内のリーフ スイッチ2個に対してデュアルホームとなっている、という状況を考えます。この場合、リーフ スイッチごとのダウンストリーム帯域幅は次のように計算されます。

$$20 \text{ servers} * 25 \frac{\text{gigabits}}{\text{server}} = 500 \text{ gigabits}$$

各リーフ スイッチのダウンストリーム帯域幅要件は、500Gbです。ただし、一部のトラフィックはリーフ スイッチのペアに対してローカルになるため、スパイン スイッチをトラバースする必要はありません。

ラックのリーフ スイッチに対してローカルなトラフィックの量は、設定でのラックの数によって決まります。ラックが2つある場合、トラフィックの50%がローカルになる可能性が高くなります。ラックが3つある場合、トラフィックの33%がローカルになる可能性が高くなります。ラックが4つある場合、トラフィックの25%がローカルになる可能性が高くなります。これ以降も同様です。言い換えれば、リモートになる可能性の高いI/Oの割合は次のようになります。

$$\text{remote\_ratio} = \frac{\text{number\_of\_racks} - 1}{\text{number\_of\_racks}}$$

この例では、ラックが2つあるので、帯域幅の50%がリモートになる可能性が高くなります。

$$\text{remote\_ratio} = \frac{2 \text{ total\_racks} - 1 \text{ rack}}{2 \text{ total\_racks}} = 50\%$$

この例ではラックが2つあるとなると、帯域幅の50%がリモートになる可能性が高くなります。リモートになると想定されるトラフィックの量を、各リーフ スイッチのダウンストリーム帯域幅で乗算すると、各リーフ スイッチによる総リモート帯域幅要件が求まります。

$$\text{per\_leaf\_requirement} = 500 \text{ gigabits} * 50\% \text{ remote\_ratio} = 250 \text{ gigabits}$$

この25 GbEネットワークの例では、リーフ スイッチ間で、250Gbの帯域幅が必要です。ただし、この帯域幅はスパイン スイッチ間で分散されるため、さらなる計算が必要になります。

各リーフ スイッチから各スパイン スイッチへのアップストリーム要件を求めるには、リモートのロードがスパイン スイッチ間でバランシングされるため、リモート帯域幅要件をスパイン スイッチの数で除算します。

$$\text{per\_leaf\_to\_spine\_requirement} = \frac{\text{per\_leaf\_requirement}}{\text{number\_of\_spine\_switches}}$$

この例では、各リーフ スイッチで、スパイン スイッチのメッシュによる250ギガのリモート帯域幅を要求することが想定されています。このロードはスパイン スイッチ（2つあると仮定）間で分散されるため、リーフとスパインとの間ごとの総帯域幅は次のように計算されます。

$$\text{per\_leaf\_to\_spine\_requirement} = \frac{250 \text{ gigabits}}{2 \text{ spine switches}} = 125 \frac{\text{gigabits}}{\text{spine switch}}$$

したがって、ノンブロッキング トポロジーでは、100Gb接続2つで合計200Gbとなり、各リーフ/スパイン スイッチ間では十分な帯域幅です。代わりに、125Gb/sを40Gb接続4つの間で分けることもできます。

各リーフ スイッチから各スパイン スイッチまでに必要な帯域幅の量を決定するための方程式をまとめると、次のようになります。

$$\frac{\text{downstream\_bandwidth\_requirement} * ((\text{number\_of\_racks} - 1) / \text{number\_of\_racks})}{\text{number\_of\_spine\_switches}}$$

注意：レプリケーションが実装されているシステムでは、こうした計算が追加のバックエンド レプリケーション ストレージ トラフィックに対応している必要があります。そうすると、先の例の要件の2倍になる可能性が高くなります。リーフ スイッチに対する25Gbインターフェイスが4つ、などです。

15.7 FHRPエンジン

Cisco vPCによるアクセスのルーティングのアーキテクチャと、ノードでのIPレベルの冗長性に向けて、Dellではノードのデフォルト ゲートウェイにFHRPを使用することが推奨されます。これにより、リーフ スイッチに障害が発生した場合に、デフォルト ゲートウェイが他のリーフ スイッチへフェールオーバーすることが可能になります。FHRPエンジンは、使用するスイッチ ベンダーによって異なります。Ciscoアーキテクチャを使用する場合、HSRPが使用されます。DellスイッチにはVRRPが使用されます。

アグリゲーション スイッチ1

```
interface Vlan103
no shutdown
mtu 9216
no ip redirects
ip address 192.168.103.2/24
no ipv6 redirects
hsrp version 2
hsrp 103
authentication text <text>
preempt
priority <value>
ip 192.168.103.1
```

アグリゲーション スイッチ2

```
interface Vlan103
no shutdown
mtu 9216
no ip redirects
ip address 192.168.103.3/24
no ipv6 redirects
hsrp version 2
hsrp 103
authentication text <text>
preempt
ip 192.168.103.1
```

図 15　Cisco Nexusアグリゲーション スイッチのペアに対するFHRPエンジン設定の例。アクティブなvPCピアはFHRPプライマリーとして機能することになり、バックアップvPCピアはFHRPセカンダリーとして機能することになります。

16 VMwareに関する考慮事項

ネットワーク接続はESXiで仮想化されていますが、本書に記載している物理ネットワーク レイアウトと同じ原則が適用されます。具体的には、Dell EMC PowerFlexの担当者に相談していない限り、MDMトラフィックを送信するリンクではMLAGを回避する必要がある、ということです。

物理ネットワークは、MDMまたはSDSが実行されている仮想マシン上のネットワーク スタック、またはVMkernelのSDCによって使用されるネットワーク スタックの観点から考慮するとわかりやすいです。ゲスト レベルまたはホスト レベルのネットワーク スタックのニーズを考慮してから物理ネットワークに適用すると、仮想スイッチのレイアウトに関する意思決定のための情報が得られます。

注意：バージョン3.5では、ネイティブの非同期レプリケーションは、VMwareベースのハイパーコンバージド システムではまだサポートされていません。したがって、Linuxベースのシステムでは、この場合、IPとスループットに関する前述の考慮事項が直ちに適用されるわけではありません。しかし、ユーザーが計画を進めたい場合は、セクション7.2.3に記載されている追加的スループットに関する考慮事項を考慮する必要があります。

16.1 IPレベルの冗長性

デュアル サブネット設定を使用してネットワーク リンクの冗長性を実現する場合、別個の2つの仮想スイッチが必要になります。このことは必須です。各仮想スイッチには独自の物理アップリンク ポートが搭載されているためです。PowerFlexがハイパーコンバージド モードで実行されている場合、この設定には3つのインターフェイス（SDC用のVMkernel、SDS用のVMネットワーク、物理ネットワーク アクセス用のアップリンク）があります。PowerFlexは、このモードでのインストールをネイティブでサポートします。

16.2 LAGとMLAG

LAGまたはMLAGが使用されている場合は、分散仮想スイッチを使用する必要があります。標準の仮想スイッチはLACPをサポートしていないため、推奨されません。LAGまたはMLAGを使用すると、物理アップリンク ポートでボンディングが行われます。

vSphereプラグインを使用したPowerFlexインストールでは、LAGまたはMLAGのインストールはネイティブにはサポートされません。代わりに、PowerFlex導入前に作成し、インストール プロセス中に選択します。

SDSまたはSDSを実行しているノードがスイッチへのリンクをアグリゲートしている場合、物理アップリンク ポート上のハッシュ モードは、「ソースIPアドレスとデスティネーションIPアドレス」または「ソースIPアドレスとデスティネーションIPアドレスとTCP/UDPポート」を使用するよう設定する必要があります。

これは、必要に応じて、2番目のレベルの冗長性としてのみ使用することが推奨されます。

16.3 SDC

SDCは、PowerFlex Storage Clientを実装するESXiのカーネル ドライバーです。ESXiカーネルで実行されるため、1つ以上のVMkernelポートを使用して、他のPowerFlexコンポーネントと通信します。繰り返しますが、ネイティブIPレベルの冗長性を実装するにあたっての一般的な推奨事項は、この場合、各VMkernelポートが別個の物理ポートへマッピングされていることです。2番目のレベルの冗長性が必要な場合は、IPレベルの冗長性に加えて、LAGまたはMLAGを分散スイッチ レイヤーに実装します。

16.4 SDS

SDSは、ESXi上の仮想ストレージ アプライアンス（SVM）の一部として導入されます。ここでも、当社推奨の実装では、ネイティブのIPレベルの冗長性を使用して、各サブネットが独自の仮想スイッチと物理アップリンク ポートに割り当てられています。2番目のレベルの冗長性が必要な場合は、IPレベルの冗長性に加えて、LAGまたはMLAGを分散スイッチ レイヤーに実装します。

16.5 MDM

MDMは、ESXi上の仮想ストレージ アプライアンス（SVM）の一部として導入されます。IPレベルの冗長性を使用することが強く推奨されます。**したがって、1つのMDMで、別個の仮想スイッチを2つ以上使用する必要があります。**

17 仮想化とソフトウェアデファインド ネットワーキング

今後の更新について述べることはまだあります。そうした短い記述を重ねて、SDNのサポート全般についての誤解を解くようにしています。

17.1 Cisco ACI

Cisco ACIによるPowerFlexの直接的なサポートや完全サポートはありません。特に、Cisco ACI上ではバックエンド ストレージ トラフィックをサポートしていません。一方、当社ではこれを、フロントエンドのお客様のトラフィックがACIファブリック上にフローするデュアル ネットワーク拡張でサポートします。

17.2 Cisco NX-OS

当社ではVxLAN EVPNリーフ スパイン ファブリックをNX-OSスタンドアロン ソフトウェアでサポートしています。

18 検証方法

18.1 PowerFlexのネイティブ ツール

ネットワーク パフォーマンスを監視するビルトイン ツールは、主に2つです。

1. SDSネットワーク テスト
2. SDS Network Latency Meter Test

18.1.1 SDSネットワーク テスト

SDSネットワーク テスト「`start_sds_network_test`」の使用方法については、『Dell EMC PowerFlex v3.5 CLI Reference Guide』を参照してください。 実行後の結果を取得するには、「`query_sds_network_test_results`」コマンドを使用します。

オプション`parallel_messages`と`network_test_size_gb`は、テストが実行されているリンク上の最大ネットワーク帯域幅よりも2倍以上大きくなるよう設定する必要があることに注意することが重要です。例：1つの10GbE NIC = 1250メガバイト * 2 = 2500メガバイト、または切り上げて3ギガバイト。この場合、コマンドではパラメーター「`--network_test_size_gb 3`」を使用する必要があります。これにより、ネットワークには十分な帯域幅が送信され、一貫性のあるテスト結果が得られるようになります。25 GbEネットワーク設定の場合、1つの25 GbE NIC = 3125メガバイト * 2 = 6250メガバイト、または6ギガバイトです。この場合、コマンドには「`--network_test_size_gb 6`」を含める必要があります。

パラレル メッセージ サイズは、システム内のコアの総数と同じにする必要があります。最大設定は16です。

注意：このテストは、各SDSで、設定されているSDSネットワークごとに実行する必要があります。

出力例：

```
scli --start_sds_network_test --sds_ip 10.248.0.23 --network_test_size_gb 8 --parallel_messages 8
ネットワーク テストが正常に開始されました。

scli --query_sds_network_test_results --sds_ip 10.248.0.23SDS with IP
10.248.0.23 returned information on 7 SDSs
    SDS 6bfc235100000000 10.248.0.24 bandwidth 2.4 GB (2474 MB) per-second
    SDS 6bfc235200000001 10.248.0.25 bandwidth 3.5 GB (3592 MB) per-second
    SDS 6bfc235400000003 10.248.0.26 bandwidth 2.5 GB (2592 MB) per-second
    SDS 6bfc235500000004 10.248.0.28 bandwidth 3.0 GB (3045 MB) per-second
    SDS 6bfc235600000005 10.248.0.30 bandwidth 3.2 GB (3316 MB) per-second
    SDS 6bfc235700000006 10.248.0.27 bandwidth 3.0 GB (3056 MB) per-second
    SDS 6bfc235800000007 10.248.0.29 bandwidth 2.6 GB (2617 MB) per-second
```

前述の例では、テストしているSDSからネットワーク セグメント上の他のすべてのSDSまでのネットワーク パフォーマンスを確認できます。1秒あたりの速度が、ネットワーク設定で予想しているパフォーマンスに近づくようにします。

18.1.2 SDS Network Latency Meter Test

「`query_network_latency_meters`」コマンドを使用して、SDSコンポーネント間での平均ネットワーク レイテンシーを表示します。優れた書き込みパフォーマンスを実現するには、SDSコンポーネント間が低レイテンシーであることが重要です。このテストを実行する際、10Gb以上のネットワーク接続が使用されている場合は、数百マイクロ秒以上の外れ値とレイテンシーを探します。

注意：これは、各SDSから、かつSDSネットワーク上で実行する必要があります。

出力例：

```
scli --query_network_latency_meters --sds_ip 10.248.0.23
SDS with IP 10.248.0.23 returned information on 7 SDSs

SDS 10.248.0.24
  Average IO size: 8.0 KB (8192 Bytes)
  Average latency (micro seconds): 231

SDS 10.248.0.25
  Average IO size: 40.0 KB (40960 Bytes)
  Average latency (micro seconds): 368

SDS 10.248.0.26
  Average IO size: 38.0 KB (38912 Bytes)
  Average latency (micro seconds): 315

SDS 10.248.0.28
  Average IO size: 5.0 KB (5120 Bytes)
  Average latency (micro seconds): 250

SDS 10.248.0.30
  Average IO size: 1.0 KB (1024 Bytes)
  Average latency (micro seconds): 211

SDS 10.248.0.27
  Average IO size: 9.0 KB (9216 Bytes)
  Average latency (micro seconds): 252

SDS 10.248.0.29
  Average IO size: 66.0 KB (67584 Bytes)
  Average latency (micro seconds): 418
```

18.2 Iperf、NetPerf、Tracepath

注意：PowerFlexを設定する前に、IperfとNetPerfを使用してネットワークを検証する必要があります。IperfまたはNetPerfに関する問題が特定された場合、調査が必要なネットワークの問題が発生している可能性があります。Iperf/NetPerfの問題が表示されない場合は、PowerFlex内部検証ツールを使用して、さらに正確な追加的検証を行います。

Iperfはトラフィック生成ツールであり、IPネットワークで可能な最大帯域幅を測定するために使用します。Iperf機能セットがあると、さまざまなパラメーターや、帯域幅やロスなどの測定値のレポートをチューニングできます。Iperfを使用する場合は、複数の並列クライアント スレッドで実行する必要があります。IPソケットあたりスレッド8個という選択が適切です。

NetPerfは、多種多様なタイプのネットワーキングのパフォーマンスを測定するために使用できるベンチマークです。これにより、一方向のスループットとエンドツーエンドのレイテンシーの両方についてのテストが可能になります。

Linuxの「`tracepath`」コマンドは、MTUサイズをパスに沿って検出するのに使用できます。

18.3 ネットワーク監視

ネットワークの正常性を監視して、ネットワークが最適な容量で動作することを妨げている問題を特定し、ネットワークパフォーマンスが低下しないよう保護することが重要です。市場で入手可能で、さまざまな機能セットを提供するネットワーク監視ツールは、数多くあります。

デル・テクノロジーズでは、次の領域を監視することを推奨します。

- 入出力トラフィック
- エラー、破棄、オーバーラン
- 物理ポート ステータス

18.4 ネットワーク トラブルシューティングの基礎

- pingを使用して、SDSとSDCとの間の接続をエンドツーエンドで確認
- コンポーネント間の接続を双方向でテスト
- SDSとMDMでは、通信がネットワークのみのラウンドトリップ時間1ミリ秒を超えないようにしてください。
- pingを使用して、コンポーネント間のラウンドトリップ レイテンシーを検証
- ポート エラー、破棄、オーバーランがないかどうかをスイッチ側で確認
- PowerFlexノードが稼働しているか検証

- PowerFlexプロセスがすべてのノードにインストールされ実行されているか検証
- 特にジャンボ フレームを使用している場合、すべてのスイッチとサーバーにわたってMTUを確認
- サイト間のSDR通信のMTUがWANにとって十分であるか検証
- スタティック ルーティング設定をサイト間のSDR通信で検証、WAN上のエンドツーエンド接続をテスト
- 可能な場合は、10Gb Ethernetに代えて、25Gb以上のEthernetにする
- OSイベントログにNICエラー、高レート（> 2%）のNICオーバーラン、ドロップされたパケットがないかどうかを確認
- 有効なNICの関連付けがないIPアドレスがないかどうか確認
- PowerFlexで必要とするネットワーク ポートがネットワークやノードによってブロックされていないか検証
- イベント ログまたはOSネットワーク コマンドを使用してPowerFlexを実行しているOSでパケットロスがないかどうか確認
- ノード上で実行されている他のアプリケーションが、PowerFlexで必要とされるTCPポートの使用を試みているか検証
- すべてのNICをフル デュープレックスに、自動ネゴシエーションをオンに、ネットワークでサポートされる最大速度を設定
- PowerFlexネイティブ ツール テストの出力を確認
- RAIDコントローラーに誤った設定がないかどうか確認（これはネットワーク関連ではなく、パフォーマンス一般の問題）
- 問題が発生した場合は、ログが上書きされる前に、可能な限りすぐに収集
- この他のトラブルシューティング、ログ収集情報、FAQについては、「Troubleshoot and Maintain Dell EMC PowerFlex v3.5」と「PowerFlex v3.5 Log Collection Technical Notes」を参照

19 まとめ

堅牢で持続可能なネットワークの設計では、選択した導入オプション、ネットワーク トポロジー、パフォーマンス要件、Ethernet、動的IPルーティング、検証方法、すべてを考慮します。Dell EMC PowerFlexクラスターは、さまざまなノード タイプ、ストレージ メディア、導入設定を含めて最大1024ノードまで拡張できます。そのため、ネットワークのインストールは将来の成長に合わせてサイジングする必要があります。PowerFlexはコンピューティングとストレージが同一のノード セット上に存在するハイパーコンバージド モードでも、ストレージ リソースとコンピューティング リソースが分離される2層モードでも導入できる、という事実もまた、意思決定に影響を与えます。優れたパフォーマンス、拡張性、柔軟性を実現するため、ネットワークはビジネス ニーズを考慮して設計される必要があります。このガイドに記載されている原則と推奨事項に従うことで、耐障害性、大規模な拡張性があり、ハイパフォーマンスのブロック ストレージ インフラストラクチャを実現できます。