HX データ プラットフォーム ストレッチ クラスタの運用

バージョン 1.1

2018 年 8 月

文書情報

文書概要	対象:	作成:
v1.0 HX 3.0 向けのストレッチ クラスタに関する情報	Cisco Field	Aaron Kapacinskas

使用目的と対象者

このドキュメントには、シスコ独自の機密情報が含まれています。ここに含まれる資料、アイデア、および概念は、Cisco® ソフトウェア ソリューションの構成にのみ使用できるものとします。

法律上の通知

このドキュメントに記載されているすべての情報は、非公開で提供されており、シスコの書面による許可なしに、一部か全体かを問わず、いかなる第三者にも公開または開示することはできません。

目次

前提条件

構成を進める前に、HX データ プラットフォームのリリース ノート、インストレーション ガイド、およびユーザ ガイドを確認することをお勧めします。インストレーション ガイドで説明されているようにデータ プラットフォームがインストールされ、機能している必要があります。ご不明な点がありましたら、シスコサポートまたはシスコ担当者にお問い合わせください。

はじめに

このドキュメントは、Cisco HyperFlex ストレッチ クラスタの管理ガイダンスを補足する運用ガイダンスの提供を目的としています。これにより、Cisco HyperFlex ユーザは、さまざまな障害シナリオに対応する復元力など、ストレッチ クラスタおよび Day-2 運用機能の特性を確認できます。完全に理解するためには、このソリューションのアーキテクチャやコンポーネントに関する専門知識が必要となります。そのため、このドキュメントでは、まず、通常クラスタとストレッチ クラスタの両方に使用される一般的な Cisco HyperFlex コンポーネントの概要について説明します。

また、ストレッチ クラスタの導入に特に関連する情報として、HX データ プラットフォーム ソリューションについて推奨される構成時の設定や導入アーキテクチャを示しています。これは製品ドキュメント（マニュアルやインストール ガイドなど）とともに使用するよう意図されています。製品ドキュメントについては、シスコ担当者にお問い合せください。

HX データ プラットフォームの概要：コンポーネントと環境

Cisco HyperFlex のストレッチ クラスタと通常クラスタは共通のアーキテクチャ コンポーネントに基づいていますが、ストレッチ クラスタには Cisco Unified Computing System™（Cisco UCS®）ドメイン、インストール プロセス、障害モードに関連する若干の違いがあります。このセクションでは、HX データ プラットフォームのコンポーネントについて簡単に確認します。

Cisco HyperFlex System は、エンドツーエンドのソフトウェア定義型インフラストラクチャ/ハイパーコンバージド ソリューションを目指しており、従来製品における妥協をすべて排除しています。Cisco HyperFlex System では、Cisco UCS サーバで構成されたソフトウェア定義型コンピューティング、強力な HX データ プラットフォーム ソフトウェアを利用したソフトウェア定義型ストレージ、そしてシスコ アプリケーション セントリック インフラストラクチャ（Cisco ACI™）ソリューションとスムーズに統合するシスコ ユニファイド ファブリックによるソフトウェア定義型ネットワーキング（SDN）が 1 つのシステムになっています。Cisco HyperFlex System では、ハイブリッドまたはオールフラッシュのストレージ構成、自己暗号化ドライブのオプション、および選択された管理ツールにより、事前統合済みクラスタを導入して、1 時間以内に稼働を開始できます。Cisco UCS サーバをコンピューティング専用ノードとして統合する機能により、コンピューティングおよびストレージ リソースを個別に拡張して、アプリケーションのニーズに的確に対応できます。

以下では、Cisco UCS、ファブリック インターコネクト、Cisco HyperFlex HX シリーズ ノードなど、このソリューションの各コンポーネントについて説明します。これらのコンポーネントは、ストレッチ クラスタでも従来のクラスタでも同じです。

Cisco Unified Computing System

物理 HX シリーズ ノードは、ハイブリッド構成かオールフラッシュ構成の Cisco UCS 220 または 240 プラットフォーム上に展開されます。

サービス プロファイルは、サーバとその LAN および SAN 接続をソフトウェアで定義したものです。1 つのサービス プロファイルで、1 台のサーバとそのストレージおよびネットワーク特性を定義します。サービス プロファイルは、Cisco UCS 6248UP 48 ポートまたは 6296UP 96 ポート ファブリック インターコネクトおよび 6332 または 6332-16UP ファブリック インターコネクトに格納され、特定バージョンの Cisco UCS Manager（ファブリック インターコネクトの Web インターフェイス）または API を使用した専用ソフトウェアによって管理されます。サービス プロファイルがサーバに導入されると、Cisco UCS Manager は、そのサービス プロファイルで指定された設定に一致するように、サーバ、アダプタ、ファブリック エクステンダ、ファブリック インターコネクトを自動的に設定します。このようにデバイス設定が自動化されることで、サーバ、ネットワーク インターフェイス カード（NIC）、ホスト バス アダプタ（HBA）、LAN スイッチ、および SAN スイッチの設定に必要な手作業が軽減されます。

HX シリーズ ノードのサービス プロファイルは、インストール時のクラスタ構築プロセスで作成され、ファブリック インターコネクトに接続されている適切なデバイス（部品番号および関連するハードウェアで識別）に適用されます。これらのプロファイルは、わかりやすい一意の名前を指定する必要があり、作成後は編集しないようにする必要があります。このサービス プロファイルは、Cisco HyperFlex System を安全かつ効率的に運用するために必要な設定（VLAN、MAC アドレス プール、管理 IP アドレス、QoS プロファイルなど）を使用して Cisco HyperFlex インストーラによって事前設定済みです。

ファブリック インターコネクト

Cisco UCS ファブリック インターコネクトは、Cisco UCS シャーシの接続先となるネットワーク スイッチまたはヘッド ユニットです。ファブリック インターコネクトは、Cisco UCS の中心となるコンポーネントです。Cisco UCS は、1 つのユニットとして動作する単一のプラットフォームにすべてのコンポーネントを統合することで、拡張性を向上できると同時にデータセンターの総所有コスト（TCO）を削減できます。ネットワークやストレージへのアクセスは、Cisco UCS ファブリック インターコネクトを通じて提供されます。HX シリーズの各ノードは、高可用性を提供できるように、ファブリック インターコネクトごとに 1 つの Small Form-Factor Pluggable（SFP）によってデュアル接続されます。この設計により、Cisco UCS 内のすべての仮想 NIC（vNIC）が同様にデュアル接続され、ノードの可用性が本質的に確保されます。vNIC の構成については、Cisco HyperFlex System のインストール時に自動的に行われ、変更しないようにする必要があります。

ファブリック インターコネクトのトラフィックとアーキテクチャ

ファブリック インターコネクトを通過するトラフィックには、一般に 2 つのタイプがあります。1 つは、クラスタ内トラフィック（ノード間）で、もう 1 つはクラスタ外トラフィック（クライアント マシンやレプリケーションに関連するトラフィック）です。すべてのファブリック インターコネクト構成に対する管理、アクセス、および変更は、Cisco UCS Manager を介して行われます。

Cisco UCS Manager の要件

Cisco UCS Manager は、Cisco UCS サービス プロファイル用および一般的なハードウェア管理用にファブリック インターコネクトを設定するためのインターフェイスです。インストール時、Cisco HyperFlex インストーラにより、Cisco HyperFlex System に適したバージョンの Cisco UCS Manager が展開されていること、またサポートされているバージョンのファームウェアがハードウェアで実行されていることが確認されます。インストール時にこれらのバージョンを必要に応じてアップグレードするオプションが提示されます。

シリアルオーバー LAN（SoL）機能は、VMware ESX の構成には必要ないため、導入後に無効にすることをお勧めします。また、インストール作業中に使用したデフォルト値や単純なパスワードも変更する必要があります。

仮想ネットワーク インターフェイス カード

vNIC の詳細については、次を参照してください。https://supportforums.cisco.com/document/29931/what-concept-behind-vnic-and-vhba-ucs [英語]

各仮想スイッチ（vSwitch）の vNIC については、順序があらかじめ定義されており、Cisco UCS Manager または ESX で変更しないようにする必要があります。変更を行うと（アクティブまたはスタンバイ ステータスを含む）、Cisco HyperFlex System の機能に影響する可能性があります。

East-West トラフィック

HX データ プラットフォームの通常クラスタでは、ファブリック インターコネクト上の East-West トラフィックは、HX シリーズ ノード間のネットワーキング トラフィックになります。このトラフィックはシステム固有であり、ファブリック インターコネクトから上流に位置するスイッチに送出されることはありません。このトラフィックの利点は、低遅延、低ホップ数、高帯域幅により非常に高速であるということです。また、このトラフィックがローカル システムの外に出ることはないため、外部インスペクションの対象になることもありません。

ストレッチ クラスタでは、このトラフィックがロケーション間のサイト間リンクを通過する必要があるため、サイトの各ファブリック インターコネクトから、ストレッチ レイヤ 2 アップリンク スイッチや対となるサイトに送出されます。ただし、このトラフィックは専用のストレージ VLAN で発生するため、安全性は維持されます。この仕組みについては、以下の「ストレッチ クラスタのアーキテクチャ」セクションを参照してください。

North-South トラフィック

ファブリック インターコネクト上の North-South トラフィックは、ファブリック インターコネクトから上流に位置するスイッチまたはルータに送出されるネットワーキング トラフィックです。North-South トラフィックは、外部クライアント マシンから Cisco HyperFlex でホストされている仮想マシンへのアクセス時、または Cisco HyperFlex System から外部サービス（Network Time Protocol（NTP）、vCenter、Simple Network Management Protocol（SNMP）など）へのアクセス時に発生します。このトラフィックには上流の VLAN 設定が適用される場合もあります。ストレッチ クラスタのサイト間トラフィックはサイト間リンクを移動する必要があります。そのため、North-South トラフィックは部分的に一般的なストレージ トラフィックに含まれることになります。ただし、説明の都合上、North-South トラフィックとは、通常、仮想マシンとエンド ユーザの間におけるやり取りのためにクラスタ（通常またはストレッチ）を出入りするトラフィックを指すものとします。

上流に位置するスイッチ

North-South トラフィックの管理には、上流に位置するスイッチまたはトップオブラック（ToR）スイッチが必要です。非ネイティブの VLAN に対応できるように上流に位置するスイッチを設定する必要があります。HX データ プラットフォーム インストーラは、デフォルトでは VLAN を非ネイティブとして設定します。ストレッチ クラスタでは、これは各サイトのレイヤ 2 の隣接関係を管理するスイッチになります。

VLAN

このソリューションでは、複数の VLAN を使用してトラフィックを分離します。VMware ESXi および Cisco HyperFlex 制御仮想マシンには管理 VLAN を使用します。また、ストレージ データ トラフィックおよびハイパーバイザ データ トラフィック（VMware vMotion トラフィック）も VLAN を使用します。ネットワークごとに個別のサブネットおよび VLAN を使用する必要があります。

デフォルトの VLAN である VLAN 1 を使用しないでください。特に分離レイヤ 2 設定を使用している場合はネットワーキングの問題が発生する可能性があります。別の VLAN を使用してください。

分離レイヤ 2 ネットワーク

分離レイヤ 2 ネットワークが環境に必要な場合は、次のドキュメントをよく確認してください。
https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/unified-computing/white_paper_c11-692008.html [英語]

使用例に応じて新しい vNIC を簡単に追加できます。シスコでは、構成への vNIC および仮想 HBA（vHBA）の手動による追加をサポートしています。この追加を安全に行う方法の詳細な手順については、『Cisco HyperFlex virtual server infrastructure (VSI) Cisco Validated Design（Cisco HyperFlex 仮想サーバ インフラストラクチャ（VSI）のシスコ検証済みデザイン（CVD））』を参照してください。
https://www.cisco.com/c/dam/global/ja_jp/td/docs/unified_computing/ucs/UCS_CVDs/HX171_VSI_ESXi6U2.pdf

このシスコ検証済みデザイン（CVD）に記載されている手順に従ってください。ピングループは使用しないでください。指定された受信者が正しく設定されていない可能性があるため、トラフィックのプルーニングが適切に行われない場合があり、接続の問題が発生することがあります。

Cisco HyperFlex HX シリーズ データ ノード

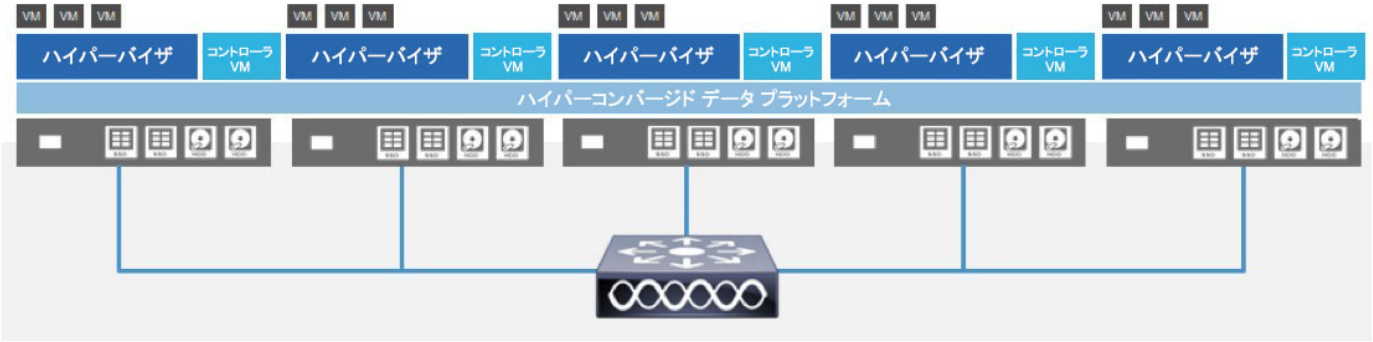
HX シリーズ ノード自体は、通常クラスタの一部であるかストレッチ クラスタの一部であるかにかかわらず、システムのハイパーバイザ用にストレージ インフラストラクチャを作成するために必要なソフトウェア コンポーネントで構成されています。このインフラストラクチャは、HX データ プラットフォームを導入する際、ノードへのインストール時に作成されます。HX データ プラットフォームでは、PCI パススルーを使用します。これによって、ハイパーバイザからのストレージ（ハードウェア）操作が不要になり、システムのパフォーマンスが向上します。HX シリーズ ノードでは、VMware インストール バンドル（VIB）という VMware 用の特別なプラグインを使用します。これは、ネットワーク ファイル システム（NFS）データ ストアのトラフィックを適切な分散リソースにリダイレクトし、スナップショットやクローニングなどの複雑な操作をハードウェアにオフロードするために使用します。

図 1 に、一般的な HX シリーズ ノードのアーキテクチャを示します。

図 1. 一般的な Cisco HyperFlex HX シリーズ ノード

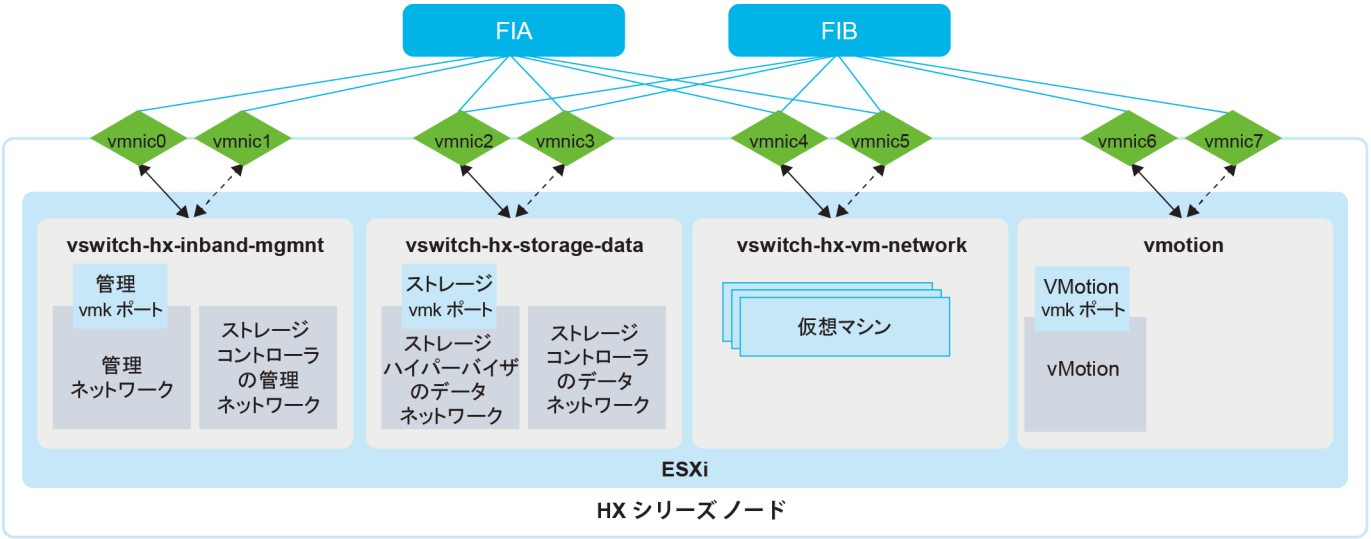
これらのノードは、Apache ZooKeeper を使用して分散クラスタに組み込まれます（図 2 を参照）。

図 2. Cisco HyperFlex 分散システム



各ノードは、仮想マシン NIC（VMNIC）および vSwitch アーキテクチャで構成されます（図 3 を参照）。

図 3. Cisco HyperFlex HX シリーズ ノードのネットワーキング アーキテクチャ



管理インターフェイス：Cisco HyperFlex Connect と VMware vCenter プラグイン

Cisco HyperFlex Connect は、クラスタ用管理ソフトで、ネイティブ HTML 5.0 ユーザ インターフェイスを持ちます。Cisco HyperFlex System 向けの vCenter プラグインが用意されており、クラスタの導入後に vCenter で利用できるもう 1 つの管理インターフェイスです。これらは別々のインターフェイスです。両方とも、Web ブラウザで HTTPS を介してアクセスでき、コマンドライン インターフェイス（CLI）および API で利用可能な同じユーザ管理（ロールベース アクセス コントロール（RBAC）を含む）の対象となります。

Apache ZooKeeper

ZooKeeper は、基本的には、階層型のキー値ストアに対する分散システム向けの集中型サービスです。大規模な分散システム向けの分散型構成サービス、同期サービス、およびネーミング レジストリを提供するために使用します。

ZooKeeper のアーキテクチャでは、冗長サービスによって高可用性がサポートされています。そのため、クライアントは、最初の呼び出しが失敗した場合には別の ZooKeeper リーダーを要求できます。ZooKeeper ノードでは、ファイル システムやツリー データ構造のように、階層型の名前空間にデータが格納されます。クライアントはノードに対して読み書きを行うことができ、これによって共有構成サービスが実現されます。ZooKeeper は、更新を包括的に管理するためのアトミック ブロードキャスト システムとみなすことができます。

ZooKeeper は、次の主な機能を提供します。

- **信頼性の高いシステム**：1 つのノードで障害が発生した場合でも適切に機能し続けるため、非常に信頼性の高いシステムとなっています。
- **シンプルなアーキテクチャ**：ZooKeeper のアーキテクチャは非常にシンプルで、プロセスの調整に役立つ共有の階層型名前空間を使用しています。
- **高速な処理**：ZooKeeper は読み取り主体のワークロードについて特に高速に動作します。
- **優れた拡張性**：ZooKeeper のパフォーマンスは、ノードを追加することで強化できます。

VMware vCenter

HX データ プラットフォームでは、VMware High Availability（HA）および Distributed Resource Scheduler（DRS）の VMware ESX クラスタリング、仮想マシンの展開、ユーザ認証、さまざまなデータ ストア操作など、クラスタ作成の特定の側面を管理するために、VMware vCenter を導入する必要があります。Cisco HyperFlex System 向けの vCenter プラグインは、vCenter 内でのシームレスな統合、およびクラスタに関する包括的な管理やレポート作成を可能にする管理ユーティリティです。

VMware ESX

ESX は、このソリューションのハイパーバイザ コンポーネントです。ゲスト仮想マシン用にノード コンピューティングやメモリ ハードウェアを抽象化します。HX データ プラットフォームは、ESX と緊密に連携してネットワークおよびストレージの仮想化を促進します。

仮想マシン

Cisco HyperFlex 環境では、セグメント化された VLAN ネットワーキングを使用して、ESX に展開されているゲスト仮想マシン用のストレージを提供します。仮想マシンは、柔軟性の高いインフラストラクチャの展開でよく見られるように、外部リソースに利用できます。

クライアント マシン

ここではクライアント マシンは、Cisco HyperFlex System に展開されたリソースにアクセスする必要がある外部ホストとして定義されます。分散アプリケーション アーキテクチャでは、このようなリソースには、エンド ユーザから他のサーバに至るまで、のあらゆる要素が含まれます。これらのクライアントは、外部ネットワークからシステムにアクセスし、ネットワークのセグメント化、ファイアウォール、およびホワイトリスト ルールによって Cisco HyperFlex の内部トラフィックから常に分離されます。

HX データ プラットフォームのストレッチ クラスタ

このセクションでは、Cisco HyperFlex ストレッチ クラスタの概要を示します。このようなクラスタを導入するビジネス上の理由についても説明します。また、このようなクラスタの物理的な制約についても説明します。

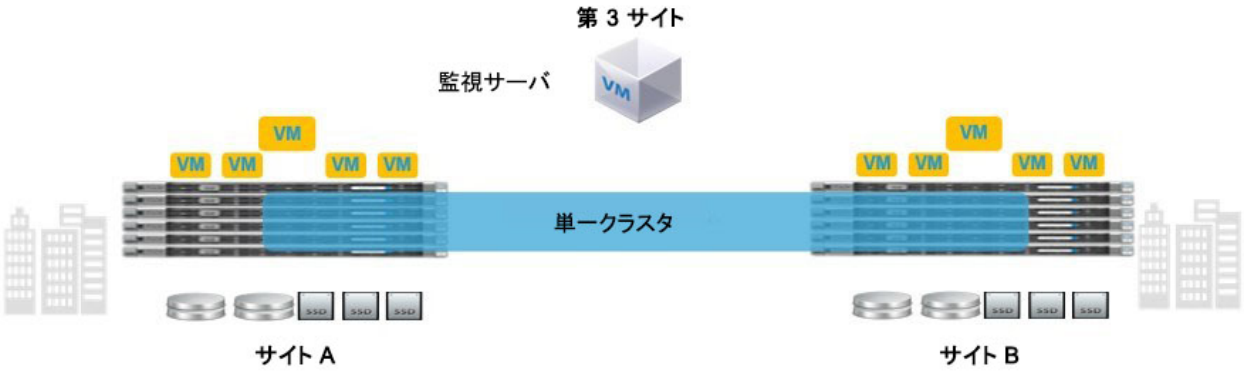
ストレッチ クラスタとは

ストレッチ クラスタは、データセンターがある場所で大きな災害が発生した場合にもビジネス継続性を提供できるよう設計されているという点で、非ストレッチ、つまり通常のクラスタとは明確に区別されます。ストレッチ クラスタは地理的冗長性を備えています。つまり、クラスタの一部は第 1 の物理的な場所に存在し、別の部分は第 2 の場所に存在します。さらに、このクラスタには、第 3 の別の場所に存在する「タイ ブレーカー」または「監視」コンポーネントが必要となります。この設計の目的は、1 つのサイトで機能が完全に失われた場合でも仮想インフラストラクチャが利用可能な状態に保たれるようにすることです。当然ながら、比較的規模の小さいさまざまな障害が発生することもあり、そのような場合にも高可用性が維持されます。これらのシナリオについては後で説明します。

よくある誤解は、ストレッチ クラスタが複数の単一クラスタをまとめたものであるとされることです。これは正しくありません。ストレッチ クラスタは、実際には、単一の分散型エンティティであり、ほとんどの状況でそのように動作します。ただし、通常のクラスタとストレッチ クラスタには異なる点がいくつかあります。この違いは、ストレッチ クラスタは展開先の地理的冗長性を提供するために特別な要件を満たさなくてはならない、という単にそれだけの理由で生じます。スプリット ブレインやノード クォーラムなど、特定の状況に適切に対応できるように、地理的冗長性にはクラスタに対する新しい要件が伴います。この点については、以下のセクションで説明します。

図 4 に、ストレッチ クラスタの主な機能を示します。

図 4. ストレッチ クラスタを構成する 3 つの主要コンポーネント



ストレッチ クラスタには注目すべき次の特性があります。

- ストレッチ クラスタは、ノードが地理的に、つまり異なる場所に分散された単一クラスタです。
- ストレージは、各サイトでローカルにミラーリングされます（タイ ブレーカーの監視エンティティは対象外）。
- サイトは、アプリケーションの書き込み要件を満たして優れたエンドユーザ エクスペリエンスを実現するために、低遅延のネットワークで接続する必要があります。
- 地理的なフェールオーバー（仮想マシン）は、通常クラスタでのフェールオーバーに似ています。
- 1 つのサイトでのノード障害は、通常クラスタでのノード障害に似ています。
- スプリット ブレインとは、どちらのサイトのノードでも互いを認識できない状況です。この状況では、ノード クォーラムを決定できない（そのために仮想マシンが実行場所を特定できない）場合に問題が発生する可能性があります。スプリット ブレインの原因は次のとおりです。
  - ネットワーク障害
  - サイト障害
- ストレッチ クラスタには監視機能があります。これは第 3 のサイトでホストされているエンティティで、スプリット ブレインの状況が発生した場合にどのサイトがプライマリになるかを決定する役割を担います。

ストレッチ クラスタに対するビジネス ニーズ

企業では、重大なインシデントや災害が発生した場合にビジネスの継続性を確保し、短期間のうちに適切に通常の運用を再開できるように計画して準備する必要があります。ビジネスの継続性とは、災害の最中も、災害後も、重要な機能を維持する組織の能力です。これには主な要素が 3 つあります。

- **復元力**：重要なビジネス機能や基盤となるインフラストラクチャについては、関連する機能の中断による影響を実質的に受けない（たとえば、冗長性や予備の容量を利用して対応する）ように設計する必要があります。
- **回復力**：重要なビジネス機能および重要度の低いビジネス機能については、特定の理由で障害が発生した場合に回復できるように対策を講じておく必要があります。
- **不測の事態への対応力**：組織では、予期しない、そして多くの場合は予測できない事象を含め、どのような重大なインシデントや災害が発生した場合にも効果的に対処できるように、汎用的な能力を高めて準備しておく必要があります。不測の事態への備えは、復元力や回復力の対策では実際には不十分だと判明した場合に最後の対応手段となります。

ストレッチ クラスタの物理的な制約

一部のアプリケーション、特にデータベースでは、書き込み遅延が 20 ミリ秒（ms）未満であることが求められます。その他の多くのアプリケーションでは、そのアプリケーションに関する問題を回避するために遅延が 10 ms 未満であることが必要とされます。このような要件を満たすには、ストレッチ クラスタ内のサイト間におけるストレッチ リンクのラウンドトリップ時間（RTT）ネットワーク遅延が 5 ms 未満であることが必要です。ストレッチ クラスタの推奨される最大サイト間距離である 100 km（約 62 マイル）では、光速（3e8 m/s）でも約 1 ms の遅延がそれだけで発生します。さらに、コード パスおよびリンク ホップ（ノードからファブリック インターコネクトを経由してスイッチまで）に一定の時間が必要であり、これも同様に、推奨される最大サイト間距離を決定する要素となります。

以下では、Cisco HyperFlex の一般的な展開に含まれるコンポーネントについて説明します。セキュアな環境を実現するには、さまざまな部分を必要に応じて強化する必要があることに注意してください。

ソリューションのコンポーネント

従来の Cisco HyperFlex の単一クラスタは、相互に接続された Cisco UCS の HX シリーズ ノードと、ファブリック インターコネクトのペアを介した上流に位置するスイッチで構成されます。1 つのファブリック インターコネクトのペアに 1 つ以上のクラスタが含まれます。ストレッチ クラスタでは、2 つの独立した（サイトごとに 1 つの） Cisco UCS ドメインが必要です。したがって、1 つのストレッチ クラスタに合計 4 つのファブリック インターコネクト（2 つのペア）が必要となります。他のクラスタは同じファブリック インターコネクトを共有できます。

図 5 および図 6 は、このような展開の一般的な物理レイアウトを示しています。図 5 は、単一サイトとそのケーブル接続および独立した Cisco UCS ドメインを示しています。図 6 は、ストレッチ クラスタのサイト A およびサイト B のラックと、各サイトのファブリック インターコネクトおよび上流に位置するスイッチを示しています。これは、8 ノード（4+4）のストレッチ クラスタで、Cisco HyperFlex HX220c ノードが各場所で使用されています。

図 5. ストレッチ クラスタ展開のサイト A：単一サイトのラックを示しており、このサイトには、HX220c M5 ノードが 4 つ、ファブリック インターコネクトが 2 つ、さらにサイト B に接続するストレッチ レイヤ 2 ネットワーク用のアップリンク スイッチが 1 つある

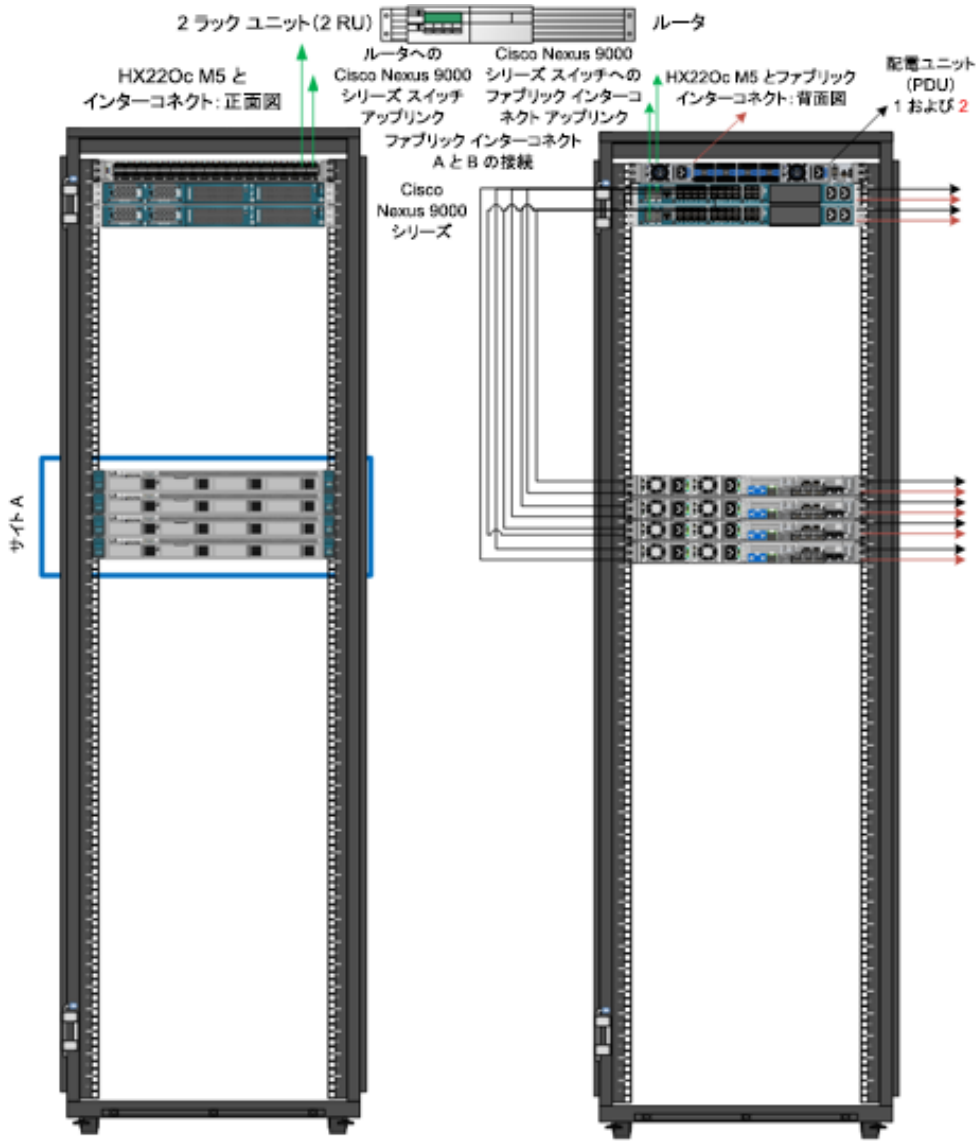
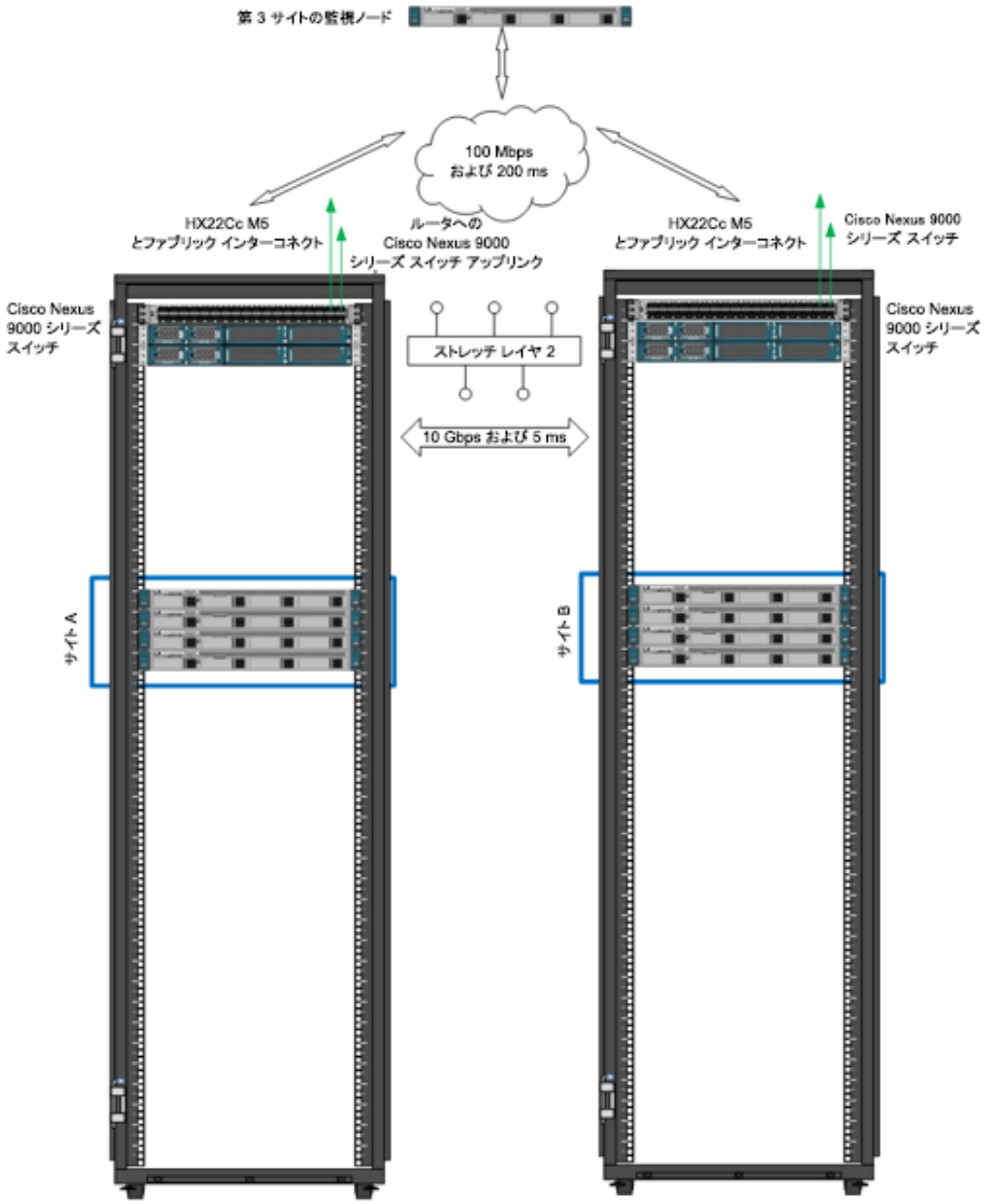


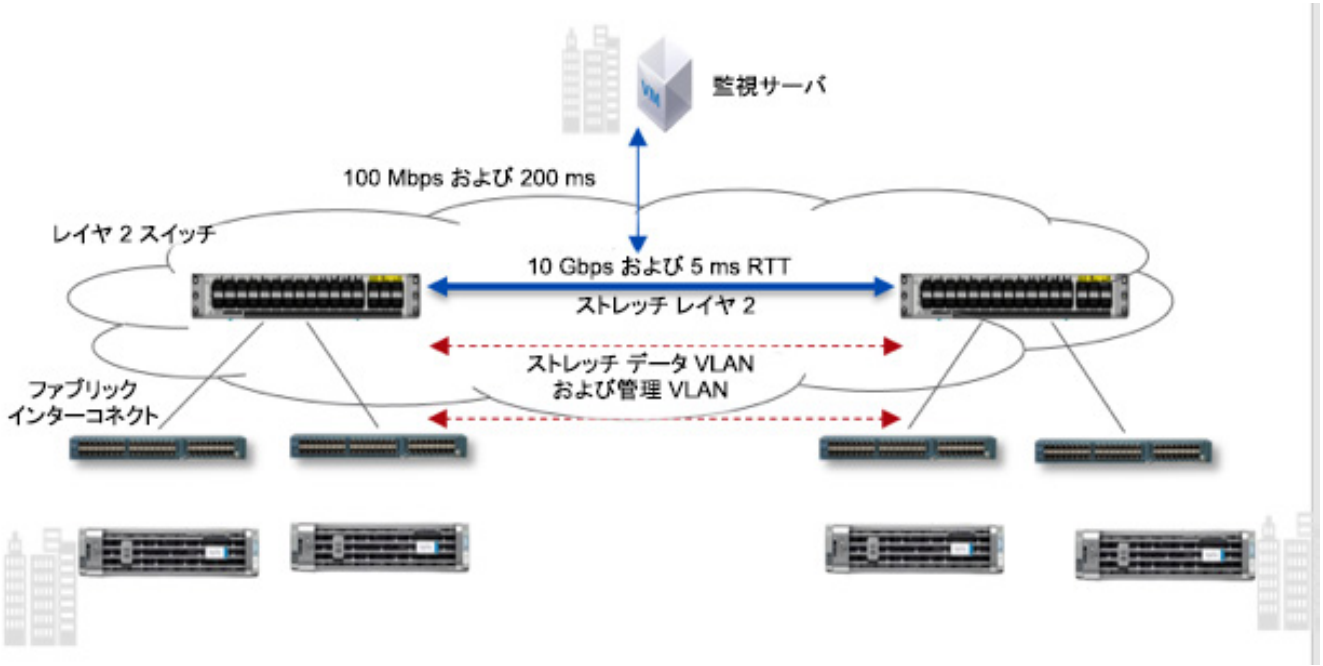
図 6. サイト A とサイト B のラック図：それぞれにファブリック インターコネクトがあるほか、ストレッチ クラスタの監視用として第 3 の論理サイトが別の場所にある



ストレッチ クラスタのアーキテクチャ

このセクションでは、ハードウェア、ネットワーキング構成、VMware の要件（ESXi および vCenter）、障害を考慮したサイジング、監視エンティティの特性など、ストレッチ クラスタの導入に関する必要事項について説明します（図 7）。VMware vSphere Enterprise Plus は必須です。これは、そのプレミアム エディションでのみ利用可能な高度な DRS 機能が Cisco HyperFlex ストレッチ クラスタには不可欠であるためです。VMware 上にストレッチ クラスタまたはメトロポリタン クラスタを実装するすべてのスタックで（ハイパーコンバージド インフラストラクチャ（HCI）以外の場合や従来型ストレージの場合でも）要件は同じです。

図 7. 一般的なストレッチ クラスタのネットワーク



ストレッチ クラスタを導入する際の最初の考慮事項は、適切なサイト間ネットワークを構築することです。ストレッチ クラスタでは、リンクに関して最低でも 10 ギガビット イーサネット接続および 5 ms の RTT 遅延が必要となります。リンクは、ストレージの通信に使用されるデータ ストレージ VLAN ネットワークのネットワーク空間に関する隣接関係を確保できるように、ストレッチ レイヤ 2 であることが必要です。サイト間のネットワークに必要とされる特性は次のとおりです。

- ストレージ データ VLAN 用として 10 Gbps（専用）
- 2 つのアクティブ サイト間で 5 ms の RTT 遅延
- ストレッチ レイヤ 2 VLAN 上のデータ VLAN と管理 VLAN
- 2 つのサイト間のストレッチ レイヤ 2 VLAN
  - ダーク ファイバと高密度波長分割多重（DWDM）レイヤ 2 および 3 の技術がサポートされています。
  - このソリューションは仮想拡張 LAN（VXLAN）には現在対応していません。
  - ストレッチ レイヤ 2 の特性
  - ストレッチ データ VLAN では、ジャンボ最大伝送ユニット（MTU）を使用する必要があります。この設定はインストーラで検証され、この設定が適用されていない場合はインストール プロセスが停止します。
  - Cisco Nexus® 5000 シリーズ スイッチは、Cisco Nexus 7000 および 9000 シリーズ スイッチと若干異なります。デフォルトのネットワーク QoS ポリシーではジャンボ MTU が許可されていませんが、スイッチ間でジャンボ スイッチ ポリシーを設定できます。
  - クラスタ内の任意の ESXi ホストから、**VMkping -I VMk1 -d -s 8972 x.x.x.x** を使用して、RTT ping をテストしてください。このチェックは、インストーラでも実行され、失敗した場合はインストール プロセスが停止します。
- アクティブ サイトと監視サイトの間で 100 Mbps の接続および 20 ms の RTT 遅延

制限

ストレッチ クラスタには、対応ハードウェアに関連する導入上の制限事項がいくつかあります。これらの制限のほとんどは、技術的な要因ではなく、単にテスト帯域幅やリリース サイクルによるものです。これらの項目について対応が行われれば、サポート対象外機能のリストから除外され、一般的な展開で利用可能になります。マイナー バージョンのリリース ノートで、サポート リストの変更について定期的に確認してください。

最小構成と最大構成の制限事項は次のとおりです。

- 最小
  - サイトあたり 2 つのファブリック インターコネクト
  - サイトあたり 2 つのノード
  - 1 つの監視エンティティ
  - 1 つの vCenter インスタンス
  - レプリケーション ファクタ：2+2
- 最大
  - サイトあたり 2 つのファブリック インターコネクト
  - サイトあたり 8 つの小型フォーム ファクタ（SFF）ノード（合計 16 個）
  - サイトあたり 4 つの大型フォーム ファクタ（LFF）ノード（合計 8 個）
  - 1 つの監視エンティティ
  - 1 つの vCenter インスタンスまたは vCenter HA インスタンス（データベースの更新ラグがない場合）
  - レプリケーション ファクタ：2+2

ストレッチ クラスタのサポートに関する制限事項（Cisco HyperFlex 3.0）は次のとおりです。

- 自己暗号化ドライブ（SED）はサポートされていません。
- コンピューティング専用ノードはサポートされていません（Cisco HyperFlex 3.5）。
- ESXi が現時点でサポートされている唯一のハイパーバイザです。VMware vSphere Release 6.0 U3 または 6.5 U1 が必要です。
- Cisco HyperFlex ネイティブ レプリケーションはサポートされていません（Cisco HyperFlex 3.5）。
- ストレッチ クラスタへの既存クラスタの拡張はサポートされていません。
- ストレッチ クラスタは新規インストールでのみサポートされています。スタンドアロン クラスタからストレッチ クラスタ構成へのアップグレードはサポートされていません。
- オンライン ローリング アップグレードは HX データ プラットフォームでのみサポートされています。Cisco UCS Manager のアップグレードはノードごとに手動で行う必要があります。
- ストレッチ クラスタは Cisco M5 ノードでのみサポートされています。M4 ノードはサポートされていません。
- 論理可用性ゾーンはストレッチ クラスタでは現在サポートされていません。

監視エンティティには第 3 のサイトに配置された ESXi が必要です（クラウド環境は現在サポートされていません）。

ファブリック インターコネクト

ストレッチ クラスタには、ファブリック インターコネクトに関する固有の要件があります。各サイトは、独立した Cisco UCS ドメイン内にある独自のファブリック インターコネクトのペアを使用して構築されます。したがって、合計 4 つのファブリック インターコネクトが必要になります。ストレッチ クラスタでは、対称型の展開が必要とされるため、各サイトに含まれるファブリック インターコネクトおよびノードの数と種類は同じであることが必要です。サイト A に 4 つのハイブリッド ノードがある場合は、サイト B にも 4 つのハイブリッド ノードがあるようにしてください。Cisco HyperFlex 3.0 では、最大クラスタ サイズは、サイトあたり 8 つのノード、つまり合計 16（8 + 8）ノードになります。

ファブリック インターコネクトおよびノード構成の詳細情報を以下に要約します。

- 合計 4 つのファブリック インターコネクトが必要です（独自の Cisco UCS ドメイン内の各サイトに 1 つのペア）。
- 1 つのドメイン内に複数のファブリック インターコネクト モデルを混在させないでください。
- ファブリック インターコネクトには、Cisco UCS Manager Release 3.2(3e) 以上が必要です。
- 既存のファブリック インターコネクトは、Cisco M5 ノードと連携できる場合にのみサポートされます。
- ノードの要件は次のとおりです。
  - 各サイトに含まれるノードの数と種類（オールフラッシュまたはオールハイブリッド）が同じであることが必要です。
  - 最大クラスタ サイズは、サイトあたり 8 つのノードです（8 + 8）。

VMware vCenter

vCenter は、通常のクラスタではきわめて重要なコンポーネントで、ストレッチ クラスタにも不可欠です。HA および DRS が構成された vCenter は、サイトで障害が発生した場合に仮想マシンの移動を自動的に管理します。優先モード（ローカルのコンピューティングおよび読み取り I/O のために仮想マシンが 1 つのサイトに固定される）で仮想マシン ホスト グループを使用することが、ストレッチ クラスタの展開で最適なパフォーマンスを実現するために必要となります。サイトのホスト グループおよび対応するアフィニティは、Cisco HyperFlex インストーラにより構築時に自動的に作成されます。

また、データ ストアでも、仮想マシン データのプライマリ コピーを検索するメカニズムとして、ホスト グループを使用して、サイトのアフィニティが維持されます。このアプローチによって、非対称 I/O メカニズムが促進されます。ストレッチ クラスタでは、このメカニズムを使用して、読み取り I/O をローカライズしながら書き込み I/O を分散する（2 つのローカルサイト コピーと 2 つのリモートサイト コピー）ことで、クラスタの応答所要時間を改善できます。ストレッチ クラスタのサイトは両方ともアクティブなため、どちらのサイトの仮想マシンについても、一方のサイトがもう一方のサイトよりもパフォーマンスについて優遇される「二流市民」タイプのシナリオによって問題が生じることはありません。

ストレッチ クラスタの展開では、1 つの vCenter インスタンスが両方のサイトに使用されます。最善のアプローチは、サイトの機能が失われた場合にも影響を受けないように、このインスタンスを第 3 の場所に配置することです。監視サイトは必ず必要なため、監視エンティティとの共存は、多くの場合、適切な選択肢となります。

vCenter インスタンスでは、ストレッチ クラスタは単一の ESXi クラスタに対応します。必ず、ストレッチ クラスタに対して HA および DRS が設定されていることを確認してください。

監視エンティティの構成

クォーラムとは、分散トランザクションが分散システムで操作の実行を許可されるために取得する必要がある最小投票数です。クォーラムベースの手法は、分散システムで一貫性のある運用を実現するために実装されます。監視ノードはこの機能を提供します。どちらのサイトも利用可能な状態であるが互いに通信できない、スプリット ブレインの状況が発生した場合には、同じ仮想マシンの 2 つのインスタンスが HA によってオンラインになることを避けるために、仮想マシン サイト リーダーを決定する必要があります。

監視エンティティは、第 3 のサイトに展開され、その場所でインフラストラクチャの ESXi 展開で使用するオープン仮想アプライアンス（OVA）ファイルとして提供されます。監視エンティティは、ZooKeeper のインスタンスを実行し（詳細については、前記の「Apache ZooKeeper」セクションを参照してください）、投票が同数に分かれた場合に決定票を投じます。

監視ノードに必要とされる特性は次のとおりです。

- 監視仮想マシンをホストするために、第 3 の独立したサイトが必要です。
- ストレッチ クラスタの各サイトに対する監視仮想マシン用の IP アドレスおよび接続が必要です。
- 監視エンティティは、ルーティング可能なレイヤ 3 ネットワーク上にあることが必要です。
- 監視ノードの最小要件は次のとおりです。
  - 仮想 CPU(vCPU):4
  - メモリ:8 GB
  - ストレージ:40 GB
  - HA:監視ノードのオプション
- 各サイトへの RTT 遅延は 200 ms 以下であることが必要です。
- 各サイトへの帯域幅は 100 Mbps 以上であることが必要です。
- Cisco HyperFlex インストーラのストレッチ クラスタ ワークフローが実行される前に、このノードを個別に展開する必要があります。

サイトと監視エンティティの間でユーザ データが送信されていないときでも、特定のストレージ クラスタ メタデータ トラフィックが監視サイトに送信されます。このトラフィックに対応するために 100 Mbps が必要であり、競合製品でも同様です。

監視エンティティは、テストに関する制限により、クラウド環境では現在サポートされていません。OVA ファイルは、テスト済みであり、ESXi プラットフォームでサポートされています。

何らかの理由により監視仮想マシンにパッチを適用する必要がある場合は、監視エンティティを一時的にオフラインにし、更新を実施して、オンラインに戻すことができます。このプロセスについては、実際の更新を実施する場合に実稼働システムを適時に再展開できるように、各段階に分けてテスト用の監視エンティティで試験することをお勧めします。この操作を実施するには、クラスタが正常な状態であることが必要です。サポートが必要な場合は、Cisco Technical Assistance Center(TAC)にお問い合わせください。

サイジング

サイジングは一般に、ワークロードを分析したり、実行する必要がある仮想マシンの要件を把握したりすることから始めます。この情報をどのように入手したかに関係なく、次は、(独力で計算したい場合を別にすれば)サイジング ツールを使用することになります。シスコでは、一般的な VSI プロファイルを使用してストレッチ クラスタのワークロード予測を実行できるサイジング ツールを提供しています。

Cisco HyperFlex サイジング ツール:https://HyperFlexsizer.cloudapps.cisco.com/ui/index.html#/scenario

ストレッチ クラスタのサイジングを行うには、データ保護に使用されるレプリケーション ファクタについて理解する必要があります。各サイトでは、レプリケーション ファクタ 2 を適用します。つまり、各サイトで 1 つのプライマリ コピーと 1 つのレプリカが保持されます。また、各サイトでは、対となるサイトにもレプリケーション ファクタ 2 を適用します。したがって、両方のサイトで、仮想マシンごとに 1 つのプライマリ コピーと 3 つのレプリカが存在することになります(レプリケーション ファクタ 4 に相当)。この構成を使用することで、各サイトは、対となるサイトの機能が失われても問題なく動作を継続でき、ローカル ディスクやノードの障害にも対応できます。

データ保護およびワークロード プロファイル(I/O 要件)について検討することで、キャパシティのニーズを満たすために必要なディスクの数と種類を特定できます。次に、vCPU および仮想マシン メモリのニーズを満たすために必要なノード数を特定する必要があります。

サイジングのガイドラインは次のとおりです。

- VSI の場合、ストレッチ クラスタを選択するためのオプションがサイジング ツールに用意されています。サイジングを行う際はこのオプションを使用してください。
- 一般に、ストレッチ クラスタでは、レプリケーション ファクタ 4 を使用します。つまり、レプリケーション ファクタ 2 + レプリケーション ファクタ 2（各サイトでレプリケーション ファクタ 2 が適用され、対となるサイトへのフル レプリケーションがあり、同様にレプリケーション ファクタ 2）ということです。この構成により、実質的にレプリケーション ファクタ 4 になります。
- 一方のサイトにレプリケーション ファクタ 2 を使用し、もう一方のサイトに対しても同じレプリケーション ファクタを適用できます。どちらのサイトからでもすべてのワークロードを実行できるようにする場合は、全体的なワークロードとしきい値を考慮に入れて、各サイトに十分なキャパシティが用意されていることを確認する必要があります。サイジング ツールでは、この検証が自動的に実行されます。
- 仮想マシンおよび vCPU のキャパシティについて検討してください。1 つのサイトですべてを適切に実行できる必要があります。
- 仮想マシンの総 vCPU キャパシティが必要です。
- 仮想マシンの総メモリ キャパシティが必要です。

障害を考慮したサイジング

通常の動作に対して展開環境のサイジングを行うだけでは不十分です。理想的には、1 つのサイトの機能が喪失し、存続しているもう 1 つのサイトがノードを失った状態になるシナリオにも対処できるように、展開環境のサイジングを行う必要があります。これは、仮想マシン ワークロード全体へのリソース配分に対応する継続運用シナリオの中でも最悪のケースです。災害が発生した場合にも真のビジネス継続性を提供するには、ストレッチ クラスタ展開の 1 つのサイトですべてを適切に実行できる必要があります。

存続しているサイトで特定の仮想マシンのみを実行するだけで十分な場合は、システムの規模を小さくすることもできますが、ディザスタ リカバリの運用手順書を策定する際は、前述の事項を理解して考慮に入れる必要があります。ストレッチ クラスタの自動リカバリ メカニズムにより、障害が発生したサイトから仮想マシンがユーザの介入なしに起動されることに注意してください。存続しているサイトのキャパシティでは対応できない場合、状況によっては仮想マシンのフェールオーバーをオフにする必要が生じることもあります。

ストレッチ クラスタの I/O パス

ストレッチ クラスタは、各サイトでアクティブ-アクティブ モードになります。つまり、各サイトで、仮想マシンごとに、プライマリ コピーと読み取りトラフィックが発生するということです。ストレッチ クラスタにはアクティブ-スタンバイ構成の概念はありません。IO Visor である Cisco HyperFlex ファイル システム プロキシ マネージャによって、どのノードでどの読み取りおよび書き込み要求を処理するかが決定されます。一般に、ストレッチ クラスタは通常のクラスタと同じように動作しますが、ホスト アフィニティや特定の障害シナリオに関する変更点があります（後述の「ストレッチ クラスタの障害モード」セクションを参照してください）。仮想マシンのアフィニティおよび 2 + 2 のレプリケーション ファクタに基づく、読み取りと書き込みのプロセスについて、以下で説明します。

読み取りパス

仮想マシン データに対するすべての読み取り操作は、ホスト グループ アフィニティを利用して、ローカルで処理されます。つまり、読み取り操作は、仮想マシンのデータ ストアが割り当てられているサイトのノードで発生します。読み取り操作は、まず、ノードのキャッシュが利用可能な場合にはそのキャッシュを使用して処理されます。キャッシュが利用可能でない場合は、永続的なディスク領域（ハイブリッド ノード内）からデータが読み取られ、エンド ユーザに提供されます。ストレッチ クラスタ内の読み取りキャッシュは、ホスト アフィニティに基づくローカル サービスを除いて、通常のハイブリッド クラスタまたはオールフラッシュ クラスタと同じように動作します。

書き込みパス

ストレッチ クラスタの書き込み操作は、読み取り操作よりも少し複雑です。これは、データの整合性を実現するために、ローカルおよびリモートのすべてのコピーがディスクに内部的にコミットされるまで、書き込み操作が仮想マシンのゲスト オペレーティング システムにコミットされたと認識されないためです。つまり、サイト A に対するアフィニティを持つ仮想マシンは、その 2 つのローカル コピーをサイト A に書き込むと同時に、その 2 つのリモート コピーをサイト B に書き込みます。繰り返しになりますが、どのノードを使用して各書き込み操作を実行するかは IO Visor によって決定されます。

Cisco HyperFlex ファイル システムは、すべてのアクティブ コピーから書き込み操作が認識されるまで無期限に待機します。したがって、書き込み操作の実行対象となっているデータのコピーをホストするノードまたはディスクが除去された場合、書き込み操作は、障害が検出されるまで（10 秒のタイムアウト値に基づく）、または検出なしに自動的に障害が修復されるまで停止します。どちらの場合でも不整合は発生しません。

サイト A の仮想マシンからの I/O 操作は、サイト A の IO Visor によって取り込まれます。サイト B の IO Visor は関与しません。書き込み I/O 操作は、データ プラットフォーム レベルでサイト B にレプリケートされます。サイト間での仮想マシンの移行が発生した場合（たとえば、サイト A から、サイト B に対するアフィニティを持つ別のデータ ストアに、VMware Storage vMotion によって移動される場合）、IO Visor で引き継ぎが行われます。仮想マシンがサイト B に移行された後は、サイト B の IO Visor によって、I/O 操作が取り込まれます。この手順は、仮想マシンのフェールオーバー プロセスの一部でもあります。仮想マシンがサイト A からサイト B に移行された後は、仮想マシンの I/O 操作は、サイト A の IO Visor ではなく、サイト B の IO Visor によって取り込まれます。

ストレッチ クラスタのインストール

インストールを行う前に、必ず、次のインストール前チェックリストを確認し、完了してください。

http://www.cisco.com/c/dam/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/HyperFlex_preinstall_checklist/Cisco_HX_Data_Platform_Preinstallation_Checklist_form.pdf [英語]

このチェックリストは、インストール プロセスを円滑かつ適時に実施するうえで不可欠です。また、インストールするバージョンの Cisco HyperFlex のリリース ノートを確認する必要もあります。

https://www.cisco.com/c/en/us/support/hyperconverged-systems/HyperFlex-hx-data-platform-software/products-release-notes-list.html [英語]

インストール プロセスには、次のような前提条件があります。

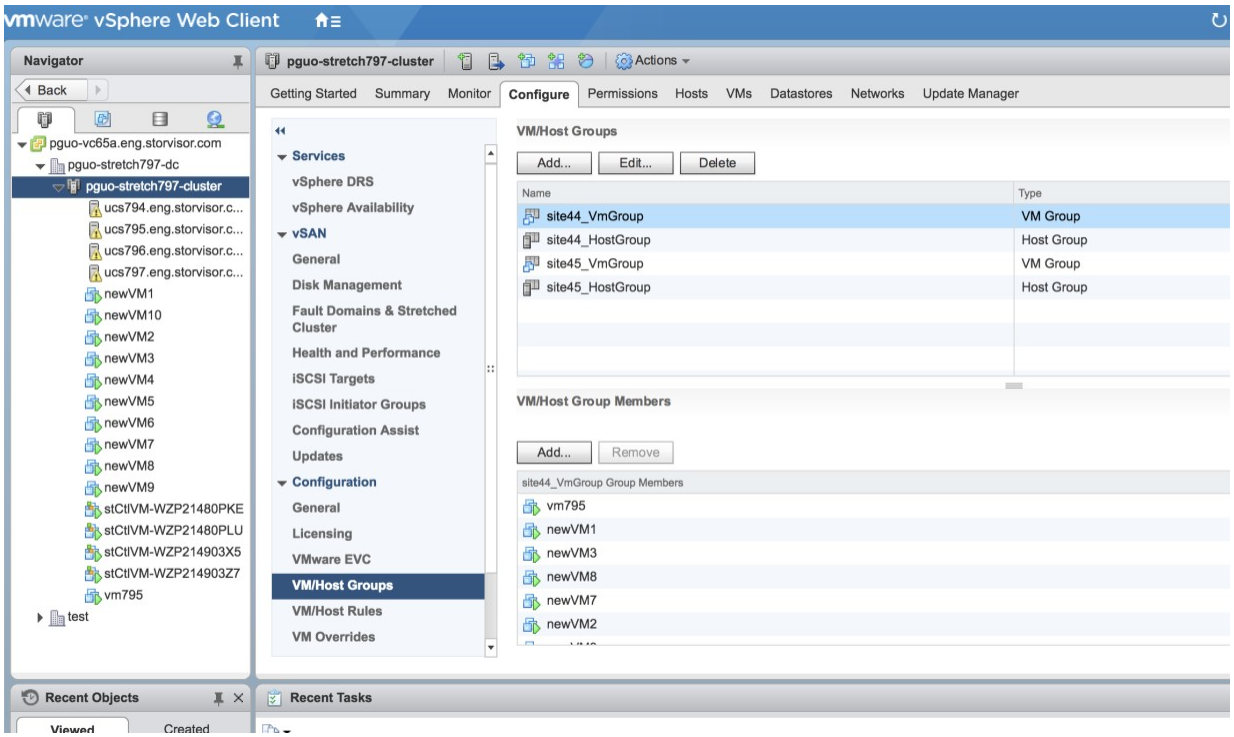
- 適切な場所を選択する必要があります。サイト間の距離は 100 km 以内でなければなりません。
- サイト間スイッチへのアップリンクを備えた 2 つのファブリック インターコネクトを各場所に配置する必要があります。
- ストレージ データ VLAN および管理 VLAN 用にストレッチ レイヤ 2 ネットワークをサイト間に配置する必要があり、ストレージ データ VLAN でジャンボ フレームがサポートされる必要があります。
- クラスタ内の任意の ESXi ホストから、**VMkping -I VMk1 -d -s 8972 x.x.x.x** を使用して、RTT ping をテストしてください。このチェックは、インストーラでも実行され、失敗した場合はインストールが停止します。
- 対称ノードを使用し、サポートされるモデルの新規インストールまたは適切な再利用を行ってください。
- 監視エンティティについては、以下の条件を満たす必要があります。
  - OVA ファイルが第 3 のサイトに配置された ESXi を使用して、展開されること。
  - 接続は 100 Mbps で、RTT 遅延が 200 ms 以下であること。
- ルーティング可能な第 3 のサイトに vCenter を事前にインストールする必要があります。
- 両方のサイトおよび監視エンティティからアクセスできるように、インストーラ OVA をネットワーク内に展開してください（必要に応じてそのインフラストラクチャ上の監視サイトに展開することもできます）。
- インストール前チェックリストを完了してください。
- リリース ノートを確認してください。

VMware vCenter にはインストール、インストール前、およびインストール後の要件もありますが、詳細なサイトの構成は自動的に処理されるため、簡単に実装できます。

- VMware DRS および HA が有効になっている必要があります。DRS は自動的に有効になります（DRS が有効になっていない場合、仮想マシンは任意のサイトで実行されます）。
- サイトごとに、サイト アフィニティを優先されるホスト グループに設定してください。仮想マシンのアフィニティ ルールおよびホスト グループは自動的に作成されます。
  - 仮想マシン アフィニティ グループとホスト アフィニティ グループが各サイトに 1 つあることを確認してください。
  - アフィニティ グループは各サイトの仮想マシンとホストで構成されます。
  - VM/Host ルールが「should」句に設定されていることを確認してください。

図 8 は、ホスト アフィニティ グループを確認できる vCenter 画面を示しています。

図 8. VMware vCenter での仮想マシンおよびホスト アフィニティ グループの確認



Cisco HyperFlex インストーラ

インストール前チェックリストを確認し、前のセクションに記載されている前提条件を満たしていれば、インストールを実行できます。初期設定プロセス中、クラスタは、Cisco HyperFlex インストーラを使用してオンサイトにインストールされます。このインストーラは、クラスタの作成後、すぐに環境から安全に削除できます。多くの場合、セキュアな環境では、インストール時に導入用のネットワークを分離します。このようなシナリオでは、設定中にインストーラを外部から利用することはできません。導入が完了した後にインストーラを削除すれば、インストーラに起因する脅威が軽減されます。

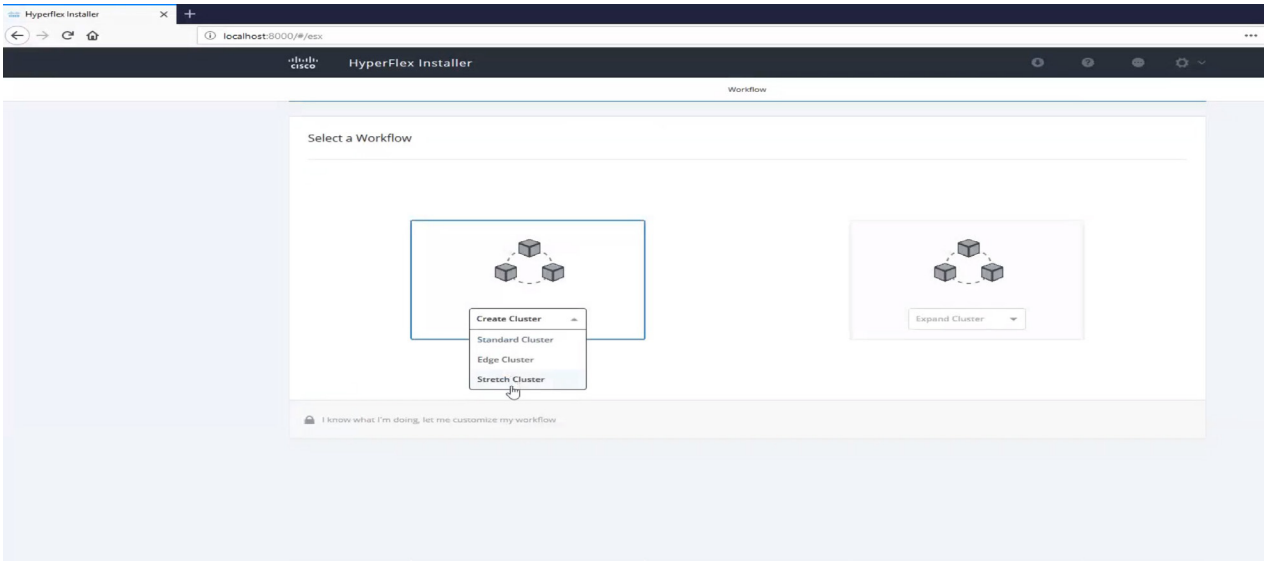
Cisco Intersight™ は、Cisco HyperFlex クラスタの作成と管理を可能にするクラウドベースのインストール、管理、およびアップグレード プラットフォームです。次の機能をサポートしています。

- デバイス コネクタを使用した、クラウドでのインストール
- デバイスの所有権
- Day-2 の運用

Cisco Intersight は、現在、通常の HyperFlex クラスタのインストールおよび ESXi への Cisco HyperFlex Edge のインストールに利用可能です。今後の Cisco HyperFlex リリースでは、Cisco Intersight をストレッチ クラスタのインストールおよび運用プラットフォームとして利用できるようになります。

インストーラを起動し、ブラウザを開いて、実行中のインストーラの IP アドレスを入力すると、ログイン画面が表示されます。アドミニストレーション ガイドに記載されているデフォルトのルート ログイン情報を使用して、開始します。ログイン後、図 9 に示されているように、ストレッチ クラスタのワークフローを選択します。

図 9. Cisco HyperFlex インストーラ：ストレッチ クラスタ ワークフローの選択



インストーラのサイト作成ワークフローを 2 回実行します（サイトごとに 1 回）。最後に、ワークフローをもう一度実行してクラスタを作成します。ワークフローを実行するたびに、インストール前チェックリストに記録したデータを入力します。このプロセスは、通常のクラスタを作成する場合と同様です。

インストーラによって、クラスタのコンポーネントが適切であり（モデル、数量など）、利用可能であることが必要に応じて確認されます。この検証プロセスにより、セキュリティやサポート性を損なう可能性がある導入上の不備がないことを確認できます。インストーラの機能は次のとおりです。

- ファームウェアとコンポート構成が条件を満たしているかを確認
- Cisco UCS Manager によってファブリック インターコネクトを調査し、適切なサーバ リストを生成
- サービス プロファイルを作成してノードに適用
  - VLAN
  - IP アドレス
  - vNIC の順序
  - QoS の設定
  - MAC アドレス プール
- 適切な VLAN およびアドレス空間を使用して ESX vSwitch を作成
- HX データ プラットフォームを展開
- ESX プラグインを展開
- クラスタを作成して展開
- ストレージ クラスタを設定して起動
- デフォルトのパスワードを設定し、ノード間通信用のセキュアな証明書を生成

インストール プロセス中、Cisco HyperFlex ユーザ インターフェイスおよび HX データ プラットフォームの設定には、強力なパスワードが適用されます。後で確認できるように、これらのパスワードを控えておいてください。

デフォルトのパスワード

インストーラを使用した導入が完了した後、デフォルトのパスワードがすべて変更または更新されていることを確認します。ESX ハイパーバイザのデフォルトのパスワードは Cisco123 です。HX データ プラットフォームのノードについては、インストール プロセス中に強力なパスワードが適用されるため、デフォルトのパスワードはありません。CLI で各 ESX ノードにログインし、必要に応じ、**passwd root** を使用してルート パスワードを更新します。

VLAN と vSwitch

VLAN は、トラフィックの種類ごと、および vSwitch ごとに作成されます。通常は、インストール プロセス中に 4 つの vSwitch が作成され、それぞれに VLAN が関連付けられます。vSwitch は、ESX 管理、Cisco HyperFlex 管理、ESX データ(vMotion トラフィック)、および Cisco HyperFlex データ(データ ストア用のノード間のストレージ トラフィック)に使用されます。vSwitch は HX データ プラットフォーム インストーラによって自動的に作成されます。

これらのスイッチが対応するゾーンは次のとおりです。

- **管理ゾーン**:このゾーンは、物理ハードウェア、ハイパーバイザ ホスト、ストレージ プラットフォーム コントローラ仮想マシン(HX データ プラットフォーム)の管理に必要な接続で構成されます。これらのインターフェイスと IP アドレスは、LAN や WAN の全体で、Cisco HyperFlex System を管理するすべての担当者が利用できる必要があります。このゾーンでは、ドメイン ネーム システム(DNS)および NTP サービスにアクセス可能であり、セキュア シェル(SSH)通信が許可されている必要があります。管理トラフィックに使用される VLAN は、Cisco UCS ドメインからネットワークのアップリンクを通過して、ファブリック インターコネクト A とファブリック インターコネクト B の両方に到達できる必要があります。このゾーンには、さまざまな物理コンポーネントと仮想コンポーネントが含まれます。
  - ファブリック インターコネクト管理ポート
  - Cisco UCS 外部管理インターフェイス。サーバやブレードによって使用され、ファブリック インターコネクト管理ポートを通じて通信します
  - ESXi ホスト管理インターフェイス
  - ストレージ コントローラの仮想マシン管理インターフェイス
  - ローミング HyperFlex クラスタ管理インターフェイス
- **仮想マシン ゾーン**:このゾーンは、Cisco HyperFlex ハイパーコンバージド システム内で実行されるゲスト仮想マシンのネットワーク I/O 操作の処理に必要な接続で構成されます。通常、このゾーンには、ネットワークのアップリンクを介して Cisco UCS ファブリック インターコネクトにトランクされ、IEEE 802.1Q VLAN ID のタグが付けられる VLAN が複数含まれています。これらのインターフェイスと IP アドレスは、LAN や WAN の全体で、Cisco HyperFlex System のゲスト仮想マシンと通信するすべての担当者やコンピュータ エンドポイントで利用できる必要があります。
- **ストレージ ゾーン**:このゾーンは、Cisco HyperFlex 分散ファイル システムにサービスを提供するために HX データ プラットフォーム ソフトウェア、ESXi ホスト、およびストレージ コントローラ仮想マシンによって使用される接続で構成されます。適切に運用するためには、これらのインターフェイスと IP アドレスが相互に通信できる必要があります。通常の運用では、このトラフィックはすべて Cisco UCS ドメイン内で発生します。ただし、一部のハードウェア障害シナリオでは、このトラフィックが Cisco UCS ドメインのネットワークのノースバウンドを通過することが必要な場合があります。そのため、Cisco HyperFlex ストレージ トラフィックに使用される VLAN は、Cisco UCS ドメインからネットワークのアップリンクを通過して、ファブリック インターコネクト A からファブリック インターコネクト B に、およびファブリック インターコネクト B からファブリック インターコネクト A に到達できる必要があります。このゾーンには、主にジャンボ フレーム トラフィックが含まれます。したがって、Cisco UCS アップリンクでジャンボ フレームが有効になっている必要があります。このゾーンには、さまざまなコンポーネントが含まれます。
  - HyperFlex クラスタ内の各 ESXi ホストの VMkernel インターフェイス(ストレージ トラフィックに使用)
  - ストレージ コントローラの仮想マシン ストレージ インターフェイス
  - ローミング HyperFlex クラスタ ストレージ インターフェイス

- **vMotion ゾーン**:このゾーンは、ホスト間でゲスト仮想マシンを vMotion によって移動できるようにするために ESXi ホストによって使用される接続で構成されます。通常の運用では、このトラフィックはすべて Cisco UCS ドメイン内で発生します。ただし、一部のハードウェア障害シナリオでは、このトラフィックが Cisco UCS ドメインのネットワークのノースバウンドを通過することが必要な場合があります。そのため、Cisco HyperFlex ストレージ トラフィックに使用される VLAN は、Cisco UCS ドメインからネットワークのアップリンクを通過して、ファブリック インターコネクト A からファブリック インターコネクト B に、およびファブリック インターコネクト B からファブリック インターコネクト A に到達できる必要があります。このトラフィックは、サイトを通過できる必要があります。

これらの vSwitch とその関連ポート グループは、高可用性を実現するために、アクティブ-スタンバイ モードの各ノードで vNIC のペアに関連付けられます。

図 10 および図 11 は、ノードの一般的なネットワーキング構成を示しています。

図 10. 仮想マシン(ユーザ)および vMotion ネットワークを示す画面

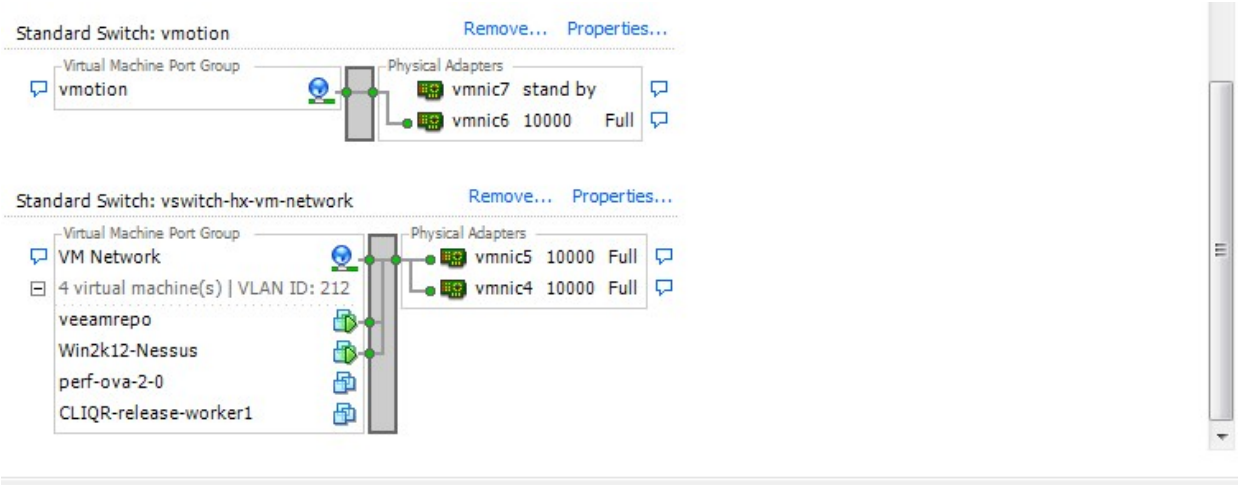
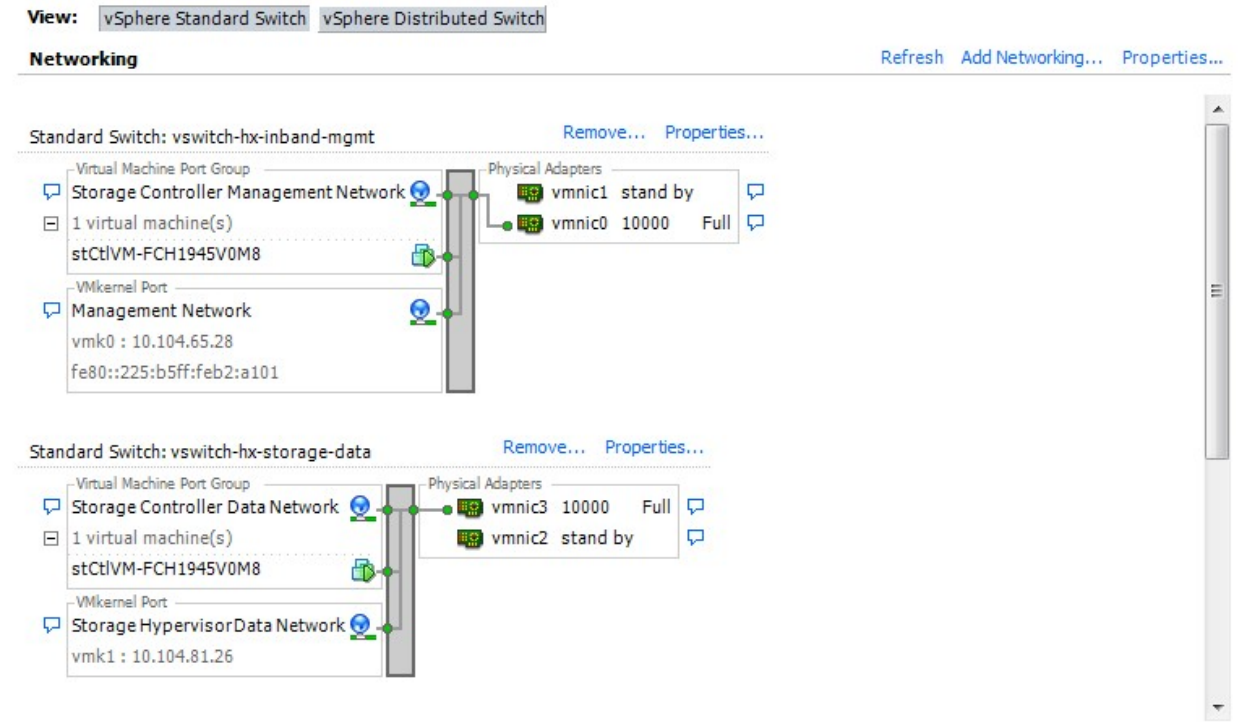


図 11. 管理および Cisco HyperFlex ストレージ データ ネットワークを示す画面



Cisco HyperFlex System での仮想分散スイッチ（VDS）の詳細については、次のリソースを参照してください。

http://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/whitepaper-c11-737724.pdf [英語]

トラブルシューティング

ストレッチ クラスタのインストールが正常に完了しない場合は、一般に、環境に関する問題が原因となっています。次の点を再確認してください。

- すべてのコンポーネントのバージョンが、インストール前チェックリストおよびリリース ノートに記載されている必要なバージョンと一致していること。
- インストール前チェックリストに示されている主要なポートがファイアウォールでブロックされていないこと。
- データ ストレージ ストレッチ VLAN でジャンボ フレームが有効になっていること。
- サイト間にストレッチ管理 VLAN があること。
- 監視エンティティが展開されており、到達可能であること。
- vCenter が展開されており、到達可能であること。

クラスタ構築が失敗したためにやり直す必要がある場合は、クラスタ導入のみの再構築と完全な再構築という 2 つの主要な再構築モードから選択できます。これらの操作のいずれを実行すべきか判断できない場合は、TAC ケースをオープンしてください。

1. クラスタ導入部分のみをやり直す
   - 監視エンティティ自体で /opt/springpath/cleanup.sh のクリーンアップ スクリプトを実行して、監視エンティティをクリーンアップしてください。
   - サポート ツールを使用してクラスタ ノードをクリーンアップすることで、部分的なクラスタの作成データをすべて破棄および削除してください。手順については、TAC にお問い合わせください。
   - クラスタ インストーラを再実行してください。

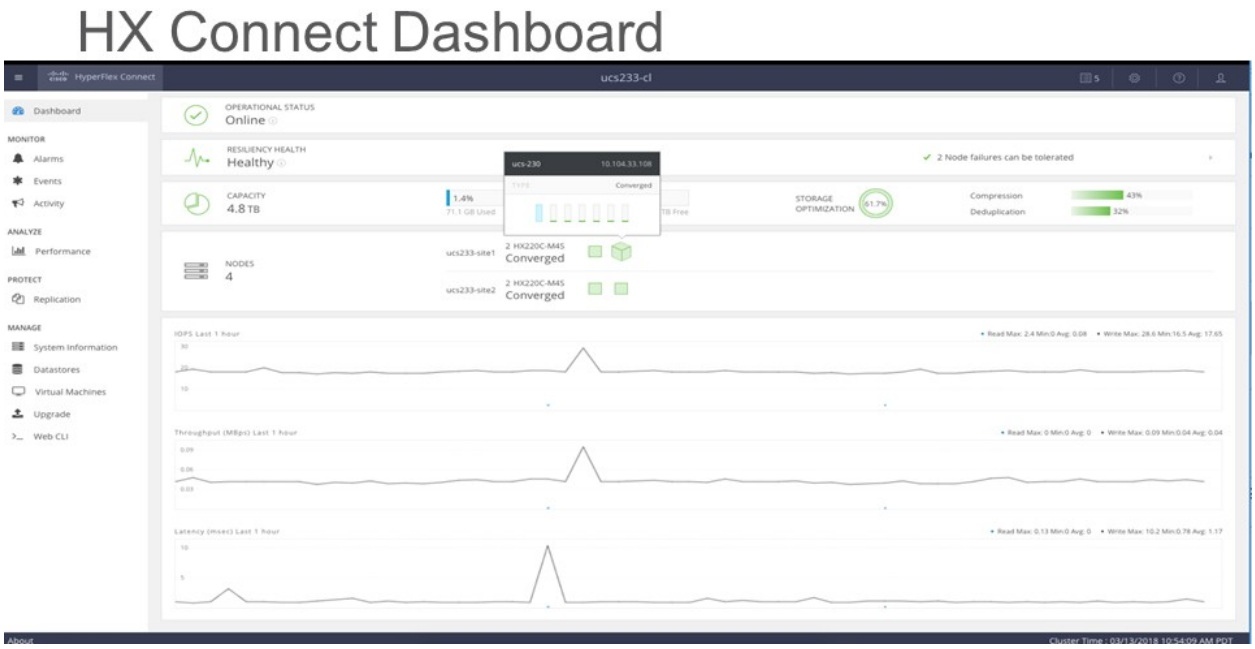
2. ESXi のインストール イメージ（CCO からダウンロード可能）と HX インストーラを使用して、最初からやり直す
   - 監視エンティティ自体で /opt/springpath/cleanup.sh のクリーンアップ スクリプトを実行して、監視エンティティをクリーンアップしてください。
   - 適切なサービス プロファイルを削除して、UCSM をクリーンアップしてください。
   - KVM 仮想ストレージを介して各ノードで UCSM に ESXi イメージをマウントし、ESXi を再インストールしてください。
   - クラスタ インストーラを再実行してください。

ストレッチ クラスタの運用

クラスタが正常にインストールされたら、データ ストアを作成して、仮想マシンを展開できます。Cisco HyperFlex Connect は、クラスタにネイティブな HTML 5 ベースのユーザ インターフェイスです。この機能にアクセスするには、インストールが完了した後にクラスタ作成の概要画面でボタンをクリックするか、（通常の場合）ブラウザでクラスタ管理 IP アドレス（CIP-M）を入力し、vCenter SSL の管理者アカウントまたはローカルの root アカウントを使用してログインします。

ログインすると、クラスタのステータスおよびパフォーマンスの概要が Cisco HyperFlex Connect ダッシュボードに表示されます（図 12）。この画面から、ヘルス ステータスの横にある矢印を使用して、ノードの数と種類、重複排除と圧縮によって節約できる総容量、パフォーマンスの概要（1 秒あたりの I/O 操作数（IOP）、スループット、遅延）、サイトベースの復元力ステータスなどを表示できます。

図 12. Cisco HyperFlex Connect ダッシュボードの表示

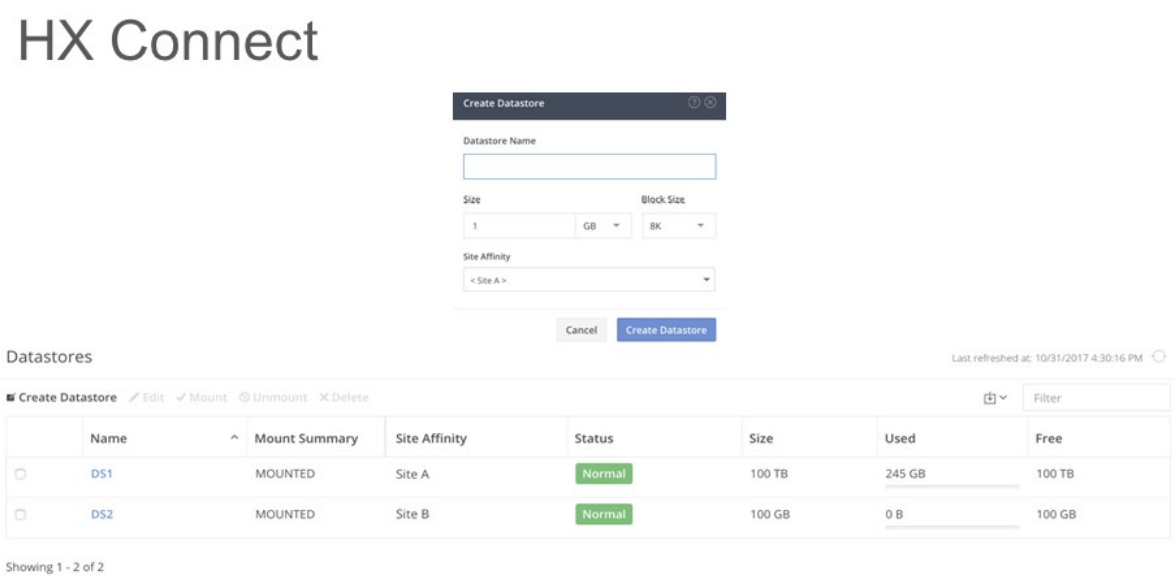


すべてではないとしても、ほとんどのクラスタ管理アクティビティに、Cisco HyperFlex Connect を使用する必要があります。特に、データ ストアの作成には、必ず Cisco HyperFlex Connect を使用してください。これにより、作成したすべてのデータ ストアについてサイト アフィニティが適切に設定されます。データ ストアを作成する場合に使用すると、以下が行われます。

- データ ストアのアフィニティが設定されます。
- 作成したデータ ストアが両方のサイトのすべてのノードにマウントされます。
- 仮想マシンが適切なサイトで起動されます。

図 13 は、Cisco HyperFlex Connect のデータ ストア作成ウィザードを示しています。サイト アフィニティの設定に注目してください。

図 13. データ ストアの作成とサイト アフィニティ



ストレッチ クラスタでの仮想マシンの展開は、vCenter を使用して通常クラスタで仮想マシンを展開する場合と同じです。

ストレッチ クラスタの障害モード

ストレッチまたはメトロポリタン（マルチサイト）単一クラスタを使用する主な理由の 1 つとして、スプリット ブレインのシナリオを回避する必要があることが挙げられます。スプリット ブレインの状況は、スコープ内で 2 つの個別のデータ セットが重複して維持されているためにデータまたは可用性の不整合が発生していることを示し、サイトの喪失、またはサーバの通信が停止してサーバ間でデータの同期が行われていないこと（サイトリンク喪失）による障害状態が原因です。監視エンティティは、このシナリオを防ぐために存在します。この点については、以下に示す障害モードの各項目で説明しています。

ストレッチ クラスタは単一クラスタであるため、多くの障害状況について、レプリケーション ファクタが 2 の単一クラスタはどのように動作するのか、という単純な疑問が生じます。サイトの機能が失われるサイト喪失（または 1 つのサイトにおける複数ノードの同時障害）が発生した場合に、1 つの場所だけを使用するクラスタの動作との違いが顕著に現れます。

ストレッチ クラスタのフェールオーバーの仕組みを理解するために、ZooKeeper について詳しく見てみましょう。アーキテクチャ上、1 つのストレッチ クラスタには ZooKeeper のインスタンスが 5 つ含まれます。各サイトに 2 つ、そして監視サーバに 1 つです。ZooKeeper の機能は、クラスタのメンバーシップおよびクラスタ全体の一貫性のあるファイル システム構成を維持することです。たとえば、各サイトに 8 つのノードがある場合（16 ノードのクラスタ）、ZooKeeper インスタンスは、各サイトの 2 つのノードでそれぞれ 1 つ（つまり、各サイトで 2 つ）、監視サーバ上でもう 1 つ実行されることになります。

障害が発生した場合に、クラスタのメンバーシップを再作成し、一貫性のあるファイル システム構成を確保するには、ZooKeeper インスタンスが少なくとも 3 つ存在する必要があります。ZooKeeper では、組み込みの投票アルゴリズム（よく知られた Paxos アルゴリズムに基づく）を使用して、この動作を実現します。

監視エンティティがダウンした場合は、1 つの ZooKeeper インスタンスが失われます。しかし、他の 4 つの ZooKeeper インスタンスは引き続き実行されているため、最低限必要な 3 つの ZooKeeper インスタンスは確保されていることになります。したがって、クラスタは影響を受けません（仮想マシンのフェールオーバーも内部での I/O の引き継ぎも発生しません）。

1 つのサイトがオフラインになった場合は、2 つの ZooKeeper インスタンスがダウンします。しかし、他の 3 つの ZooKeeper インスタンスは引き続き実行されているため、この場合も最低限必要な 3 つの ZooKeeper インスタンスは確保されていることになります。したがって、クラスタは影響を受けません。VMware HA により、仮想マシンは存続しているサイトに自動的にフェールオーバーされます。この障害は、単一クラスタで半数のノードが失われた場合と同様に処理されます。

ZooKeeper インスタンスをホストしている 1 つのノードがダウンした場合は、ZooKeeper アルゴリズムにより、別のノードが ZooKeeper 用として再選出されます。ただし、他の 4 つの ZooKeeper インスタンスは引き続き実行されているため、最低限必要な 3 つの ZooKeeper インスタンスは確保されていることになります。したがって、クラスタは影響を受けません。影響を受けた仮想マシンのみが同じサイトの存続しているノードにフェールオーバーされます（この移動はストレッチ クラスタの DRS ルールによって管理されます）。この障害は、単一クラスタで 1 つのノードが失われた場合と同様に処理されます。

障害の種類

障害の種類とその障害への対応を以下にまとめます。

- ディスク喪失
  - キャッシュ ディスク：この障害は、通常のクラスタの場合と同様に処理されます。そのサイト内の他のキャッシュ ディスクによって要求が処理され、障害が発生したコンポーネントを交換するまでキャッシュの全体的なキャパシティは減少した状態になります。
  - 永続ディスク：この障害は、通常のクラスタの場合と同様に処理されます。2 分間のタイムアウト期間が経過した後、残っているキャパシティを使用して、障害が発生したディスクのデータが再構築されます。
- ノード喪失
  - 1 つ：障害が発生したノードは、2 時間のタイムアウトが経過した後に再構築されます。または、それよりも前に、手動で再構築します。
  - 複数：複数のノード喪失が同時に発生した場合、そのサイトはオフラインになり、サイトのフェールオーバーが行われます。

- ファブリック インターコネクト喪失
  - 1 つ：障害が発生したファブリック インターコネクトが復元されるまで、そのサイトの冗長ファブリック インターコネクトによってデータが処理されます。
  - 2 つ：そのサイトはオフラインになり、サイトのフェールオーバーが行われます。
- 監視エンティティ喪失
  - 何も起こりません。クラスタは影響を受けません。監視エンティティは修復後にオンラインに戻されます。
- 監視仮想マシンの誤削除
  - リカバリ プロセスについては、Cisco TAC にお問い合わせください。
- スイッチ喪失（単一サイト）
  - 1 つ：サイトに複数のスイッチがある場合、障害が修復されるまで、冗長スイッチによってデータが処理されます。各サイトにアップリンク スイッチが 1 つしかない場合は、サイトのフェールオーバーが行われます。
  - 2 つ：そのサイトはオフラインになり、サイトのフェールオーバーが行われます。
- サイト喪失
  - そのサイトはオフラインになり、サイトのフェールオーバーが行われます。
- サイト リンク喪失
  - 2 つのサイト間のネットワークで障害が発生しただけ（ケーブルが破損した場合や、いずれかのサイトのネットワーク ポートで障害が発生した場合など）で、2 つのサイトのノードは引き続き機能しているシナリオでは、次のプロセスが実施されます。
    1. ストレッチ クラスタを作成する際、ZooKeeper リーダーを確立するために、一方のサイトにバイアスがかけられます。これは、より大きいノード ID を割り当てることによって行われます。この説明では、クォーラム サイトはサイト A とします。
    2. ネットワークの切断が発生すると、監視エンティティと、ZooKeeper リーダーがあるサイトのノードによって、サイト A でクォーラムが形成されます。
    3. もう一方のサイト（サイト B）のノードは電源がオンになった状態のままであり、このサイト（サイト B）のローカル IO Visor インスタンスからの I/O 操作では、書き込み I/O 操作を実行できなくなります。これにより、分離されたサイト（サイト B）の一貫性が確保されます。stcli cluster-info コマンドでは、これらのノードについては、物理的には電源がオンになっていても、クラスタで利用不可として表示されます。
    4. サイト A が ZooKeeper クォーラム サイトであるため、ZooKeeper の更新がサイト B で認識されるのは一定期間後（タイムアウトして障害が検出された後）になります。最終的に、サイト B の ESX 上にある IO Visor は、別のノード、つまり実際の I/O プライマリ ノード（サイト A の ZooKeeper クォーラム内に存在）と通信する必要があることを認識します。ネットワーク接続がないため、ユーザ仮想マシンがこのサイト（サイト B）上に引き続き存在していたとすると、サイト B では、その I/O 操作を再試行し続け、結局は「全パス ダウン」（APD）と認識することになります。サイト B に仮想マシンが最終的に残っていない（これらの仮想マシンは他の ESX ホストにフェールオーバーされるため）ことを手動で確認する必要があります。
    5. 仮想マシンは ZooKeeper リーダーがあるサイトにフェールオーバーされます。仮想マシンのフェールオーバーは VMware HA および DRS によって行われます。
    6. ネットワークが復元されると、フェンシング（分離）されたサイト B のノードがクラスタで再び利用可能になります。サイト間の自動再同期が行われます。ただし、仮想マシンのフェールバックは自動的には行われません。

障害対応の要約

表 1 に、前述した障害モードの要約と特定の状況に関する追加情報を示します。ここでは、別々の重大な障害が複数同時に発生する状況（両方のサイトの喪失と監視エンティティの喪失が同時に発生するなど）は考慮されていないことに注意してください。このような障害では、常に、クラスタがオフラインの状態になります。

表 1. 障害対応

コンポーネント障害	クラスタの動作	クォーラムの更新	仮想マシンの再起動	サイトの状態	クラスタの状態
単一サイトのキャッシュ ディスク	サイトはオンラインのままであり、キャッシュのキャパシティが減少します。	×	×	オンライン	オンライン
単一サイトの永続ディスク	サイトはオンラインのままであり、キャパシティが減少します。残っているキャパシティを使用して、2 分後に再構築されます。	×	×	オンライン	オンライン
両サイトのキャッシュ ディスク	サイトはオンラインのままであり、キャッシュのキャパシティが減少します。	×	×	オンライン	オンライン
両サイトの永続ディスク	別々のノードで障害が同時に発生した場合は、複数ノードの障害に該当し、サイトがオフラインになります。 別々のノードで障害が同時ではなく別々の時期に発生した場合は、クラスタは単一ディスクの障害時と同様に動作し、キャパシティが減少します。 同じノード上で障害が同時に発生した場合は、ノード障害に該当します。	○ × ×	○ × ×	オフライン オンライン オンライン	オンライン オンライン オンライン
単一サイトの単一ノード喪失	ノードは、2 時間後に、またはそれよりも前に手動で、再構築されます。	×	×	オンライン	オンライン
単一サイトの複数ノード喪失	サイトがオフラインになります。	○	○	オフライン	オンライン
単一サイトの単一ファブリック インターコネクト喪失	サイトには影響ありません。ファブリック インターコネクトを復元します。	×	×	オンライン	オンライン
単一サイトの両ファブリック インターコネクト喪失	サイトがオフラインになります。	○	○	オフライン	オンライン
両サイトの単一ファブリック インターコネクト喪失	サイトには影響ありません。ファブリック インターコネクトを復元します。	×	×	オンライン	オンライン
両サイトの両ファブリック インターコネクト喪失	両方のサイトがオフラインになります。	–	–	オフライン	オフライン
監視エンティティ喪失	サイトには影響ありません。監視エンティティを復元します。	×	×	オンライン	オンライン
単一サイトの単一スイッチ喪失	そのサイトに冗長スイッチ機能が存在する場合は、影響ありません。スイッチを復元します。 そのサイトにスイッチが 1 つしかない場合は、サイトがオフラインになります。	× ○	× ○	オンライン オフライン	オンライン
単一サイトの両スイッチ喪失	サイトがオフラインになります。	○	○	オフライン	オンライン
両サイトの単一スイッチ喪失	各サイトに冗長スイッチ機能が存在する場合は、影響ありません。スイッチを復元します。 各サイトにスイッチが 1 つしかない場合は、サイトがオフラインになります。	× –	× –	オンライン オフライン	オンライン オフライン
両サイトの両スイッチ喪失	両方のサイトがオフラインになります。	–	–	オフライン	オフライン
サイト喪失	ZooKeeper インスタンスにより、クラスタ グループに関する情報が維持され、クォーラムが形成されます。サイトが失われると、ZooKeeper の通信が消失し、サイトのフェンシングが実行されて、クラスタのクォーラムが再定義されます。ZooKeeper と DRS ルール（アフィニティ、グループなど）により、同じ仮想マシンが両方のサイトで同時に実行されないように処理されます。	○	○	オフライン	オンライン
サイト リンク喪失	前述の説明を参照してください。	○	○	オンライン	オンライン

サイトのフェールオーバーが行われた場合でも、障害が発生したサイトの仮想マシンが、存続しているサイトで起動された後は、通常どおりに運用が継続されます。仮想マシンおよび IO Visor の動作は、前述のサイト リンク喪失に関する説明で示されているとおりです。ダウンしたサイトが復元され、存続しているサイトおよび監視エンティティとの通信が再確立された後、復元されたサイトの元のデータ ストア（サイト アフィニティに基づく）に仮想マシンを移動できます。このプロセスには Storage vMotion を使用してください。これにより、元のる場所に仮想マシンが戻された後でアフィニティと適切な IO Visor のルーティングが行われます。

関連情報

追加情報については、次のリソースを参照してください。

- Cisco HyperFlex 3.0 と VSI:
  https://www.cisco.com/c/en/us/td/docs/unified_computing/ucs/UCS_CVDs/HyperFlex_30_vsi_esxi.html#_Toc514225504 [英語]
- Cisco HyperFlex サイジング ツール:
  https://HyperFlexsizer.cloudapps.cisco.com/ui/index.html#/scenario
- Cisco HyperFlex インストール前チェックリスト:
  http://www.cisco.com/c/dam/en/us/td/docs/hyperconverged_systems/HyperFlex_HX_DataPlatformSoftware/HyperFlex_preinstall_checklist/Cisco_HX_Data_Platform_Preinstallation_Checklist_form.pdf [英語]
- Cisco HyperFlex リリース ノート:
  https://www.cisco.com/c/en/us/support/hyperconverged-systems/HyperFlex-hx-data-platform-software/products-release-notes-list.html [英語]
- Cisco HyperFlex と VDS:
  http://www.cisco.com/c/dam/en/us/products/collateral/hyperconverged-infrastructure/whitepaper-c11-737724.pdf [英語]
- ZooKeeper:
  https://en.wikipedia.org/wiki/Apache_ZooKeeper [英語]
- Cisco HyperFlex と分離レイヤ 2 ネットワーク:
  https://www.cisco.com/c/en/us/solutions/collateral/data-center-virtualization/unified-computing/white_paper_c11-692008.html [英語]
- Cisco UCS での vNICS と vHBA:
  https://supportforums.cisco.com/document/29931/what-concept-behind-vnic-and-vhba-ucs [英語]

お問い合せ先

シスコシステムズ合同会社

〒107 - 6227 東京都港区赤坂9-7-1 ミッドタウン・タワー
http://www.cisco.com/jp