VMware Virtual SAN 利用時の vSphere HA 設定に関する ベストプラクティス

～Virtual SAN 障害時の動作と最適デザイン～

2014年12月

ヴイエムウェア株式会社

リードシステムズエンジニア　岡野浩史

vmware

更新履歴

バージョン	日付	内容	更新者
1.0	2014 年 12 月	初版	VMware 岡野浩史

目次

はじめに

VMware Virtual SAN は、vSphere 5.5 Update 1 以降の ESXi カーネルに組み込まれた魅力あふれるストレージの新機能で以下のような特徴を持ちます。

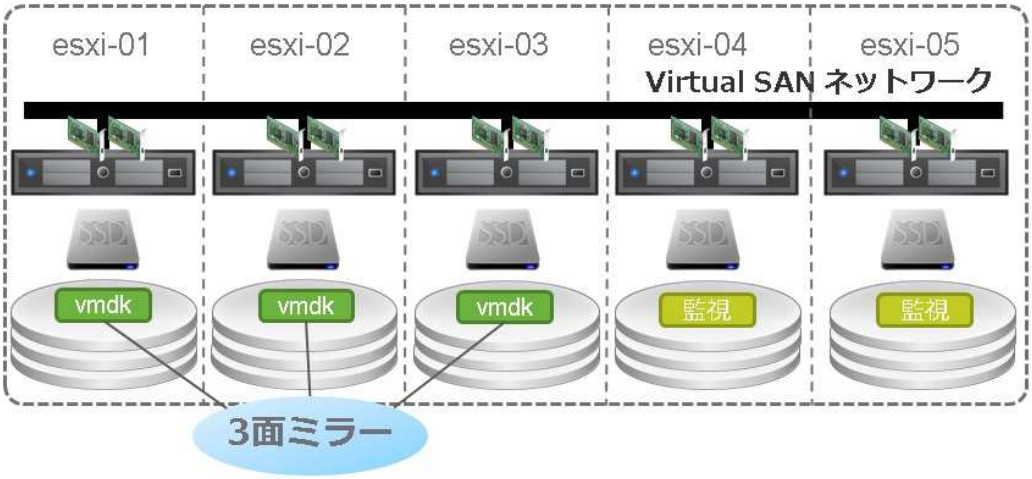
- ローカルサーバに搭載されたディスクを利用した共有ストレージ
  大容量安価な磁気ディスクと、高速低遅延なフラッシュデバイスを組み合わせたハイブリッド型
- ストレージポリシーによる管理
  可用性やパフォーマンスを仮想ディスクの粒度で定義
- 柔軟な拡張
  ホスト追加による動的なストレージ拡張（3～32 ノードをサポート）

Virtual SAN では、コンピューティング機能とストレージの機能を共にサーバで提供しますので、上記の通り非常に拡張性に富んだストレージサービスの提供が可能となるのですが、ストレージサービスをサーバに統合しているが故、主に仮想マシンのサービスレベルに関しては少しばかり注意が必要となります。本資料ではこの点にフォーカスを当ててご説明致します。なお、この資料は、vSphere 5.5 Update 2 (2014 年 9 月)の情報で作成させていただいており、将来的に変更になる可能性がありますのであらかじめご了承下さい。

Virtual SAN の耐障害性

ストレージポリシーと仮想ディスクの配置

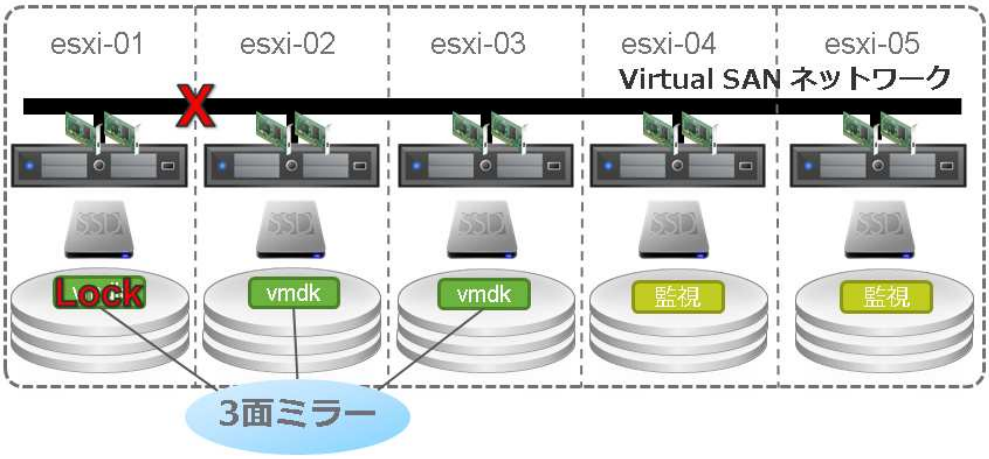
一般的な RAID コントローラでは、Disk 障害に対応するためにミラーやパリティを構成しますが、Virtual SAN の場合はホスト自体の障害に対応するため、例えば、”許容障害数＝2" のストレージポリシーで作成された仮想マシンの仮想ディスクは、3 台のホストにまたがって書き込まれます。



障害時の動き

Virtual SAN は一種のクラスタとして動作します。お互いの死活監視は Virtual SAN ネットワークを利用したハートビートのみで実装されています。Virtual SAN ネットワークはチーミングによる冗長化が可能ですしその重要性を考えると冗長化は必須とも言えますが、万が一 Virtual SAN ネットワークが切断されてしまうと、ホスト自体が障害で落ちてしまったのか単なる Virtual SAN ネットワークの障害なのか知るすべがありません。これは、ハートビートデータストアやゲートウェイの Ping を使って、相手の状態や自分自身の状態をインテリジェントに確認していく vSphere HA とは大きく異なる点なのですが、実は凄く理にかなっています。

vSphere HA の障害時の優先事項は仮想マシンの稼働を最大化させることです。このため、上記の通りインテリジェントに状態を判断し、仮想マシンへのアクションを適切にハンドリングすることを試みるのですが、Virtual SAN が、障害時に最も優先する事項は、”Disk の一貫性を保つ”ことです。



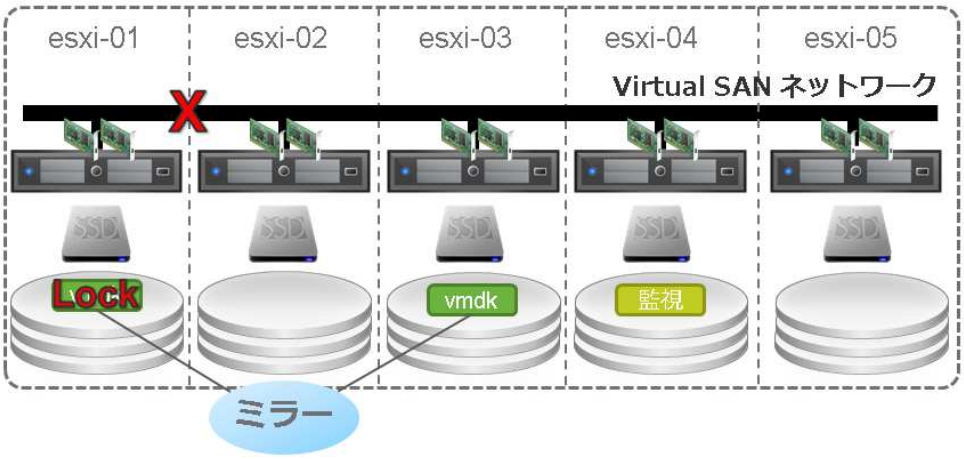
例えば、上記のような 3 面ミラーの構成時に、esxi-01 が何らかの障害により Virtual SAN ネットワークから切断されてしまった場合、分断されたミラーが共にアクセス可能な状態で放置されると、互いに異なる変更が発生し、ミラーの一貫性が失われる危険性があります。これを防ぐため、Virtual SAN は障害時、片方のミラーオブジェクトを即時にロックすることによりミラーオブジェクトの一貫性を保ちます。これが障害時に Virtual SAN が最も優先する事項です。つまり、Virtual SAN は、インテリジェントに相手の状態を把握することよりも、ミラーオブジェクトの一貫性を保つために”即時性”を最優先しているといえます。

オブジェクトロックアルゴリズム

先述の通り Virtual SAN は障害時、必要に応じ特定のディスクオブジェクトをロックします。このロックの判断は、

“ディスクオブジェクト数” +”監視オブジェクト数”

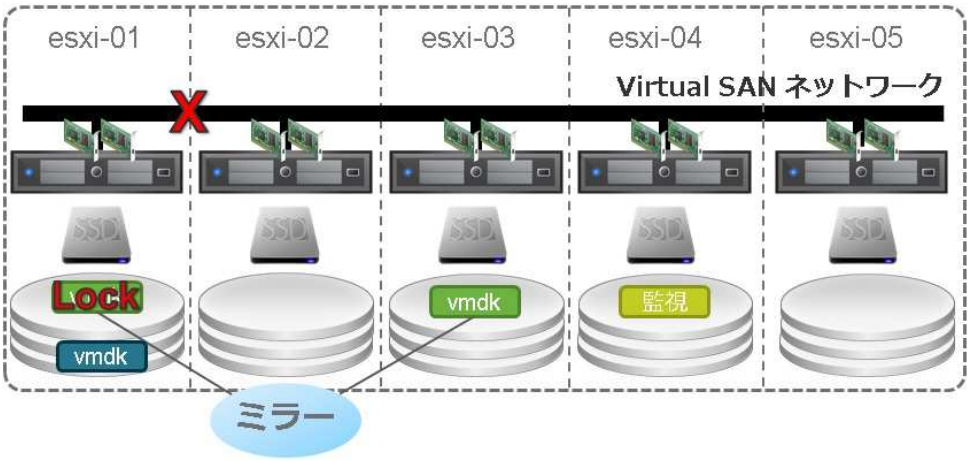
がオブジェクト総数の過半数を獲得出来たか否かよって判断されます。その結果過半数を獲得できなかったオブジェクトがロックされます。例えば、許容障害数 = 1 で仮想 Disk を作成すると、下記の様に、2 面のミラーオブジェクトとは別に”監視オブジェクト（英語名 Witness）”があらかじめ作成されます。



この状態で esxi-01 の Virtual SAN ネットワークが何かしら障害を起こしてしまった場合、それぞれのパーティションには、

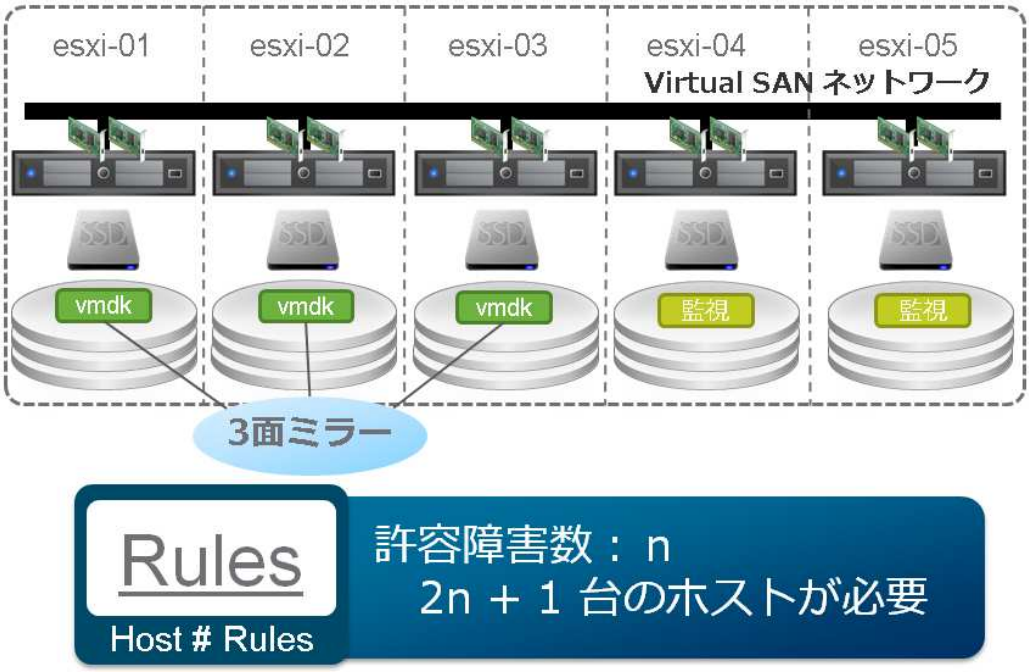
esxi-01 側・・・Disk オブジェクトが 1 つ
esxi-02～05 側・・・Disk オブジェクト 1 つ + 監視オブジェクト 1 つ = 計 2 つ

のオブジェクトが存在することになり、過半数を獲得できなかった esxi-01 上に存在する Disk オブジェクトがロックされます。一方、過半数を獲得した ESXi02～05 側の Disk オブジェクトはアクセス可能なオブジェクトとして稼働を続けます。



このオブジェクトのロックは、どちらの Virtual SAN パーティションを生かすかという全体的な判断ではなく、仮想ディスクごとに行われます。例えば、同じ環境に上の図の通り、許容

障害数 = 0 で作成されていた仮想ディスクオブジェクトが存在していた場合（下記青色の vmdk）、こちらのオブジェクトは esxi-01 上でロックされることなく稼働を続けます。



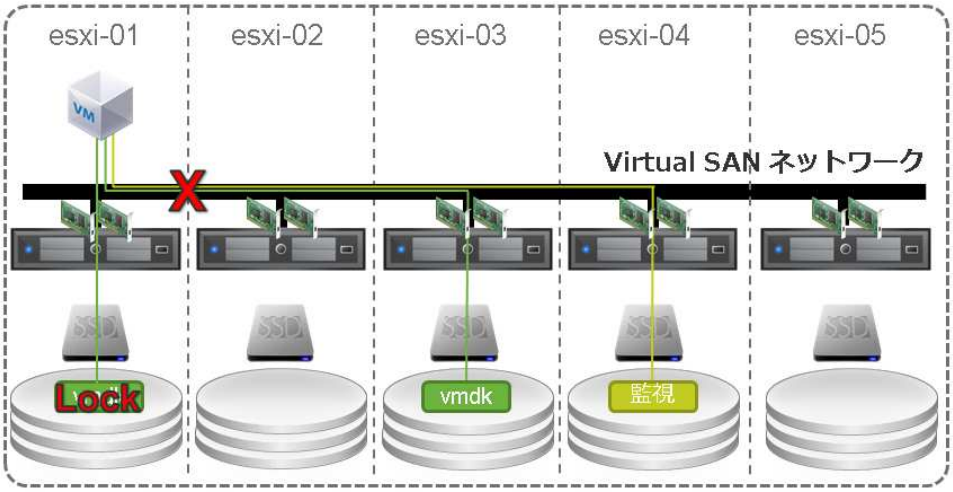
少し話はそれますが、上記した内容を鑑みると、許容障害数 = n の場合、障害時に少なくとも n+1 個のオブジェクトが正常なホスト側に残る必要があることが分かります。つまり、許容障害数 = n を構成するためには、最低でも

n + (n+1) = 2n+1 台

のホストが必要になります。例えば、3 面ミラー（許容障害数 = 2）を構成するためには、最低 5 台のホストが必要ということになります。なお、仮想マシンに関するオブジェクトは仮想ディスクファイル（vmdk ファイル）の他に、仮想マシンの構成ファイル（vmx ファイル）などが含まれる VM ホームオブジェクトも存在しますが、ロックのアルゴリズムは仮想ディスクファイルと同様となります。

仮想マシンの可用性

あらかじめ監視オブジェクトを必要数準備しておき、障害時に過半数のオブジェクトが獲得出来なかったオブジェクトをロックするというアルゴリズムは、障害対応の即時性が高く仮想ディスクのイメージの一貫性を保証する優れたオブジェクト管理の仕組みです。しかしながら、仮想マシンのサービスレベルを守るという面で考えると、ちょっと困ったことが起こる可能性があります。例えば許容障害数 = 1 で作成された仮想マシンが esxi-01 上で稼働し、仮想ディスクが以下のように配置されている場合の障害時の動きについて考えてみます。



esxi-01 がネットワーク障害で分断されると、以下のことが起こります。

- esxi-01 上のディスクオブジェクトがロックされる（過半数を確保できないため）
- ロックされていないミラーオブジェクトは esxi-03 上に存在
- esxi-01 上の仮想マシンの I/O は Virtual SAN ネットワークがアクセス不可能なため esxi-03 に到達できない

このため、仮想マシンの I/O は停止してしまいます。

この仮想マシンを正常に稼働させるには、esxi-02～05 側で仮想マシンを稼働させる必要がありますが、この動きは Virtual SAN ではなく、vSphere HA により担保されます。Virtual SAN と vSphere HA は共にクラスタに定義するサービスの 1 つで、同時に利用することも可

能です。しかしながら、例えば上記の例の様な場合に、esxi-01 側のミラーオブジェクトがロックされ、アクセス可能なミラーオブジェクトが esxi-03 にあることを Virtual SAN から vSphere HA に通知するような連係機能は現在のところ実装されていません。つまり、”esxi-02～05 側で仮想マシンを稼働させる”ためには vSphere HA が独自の判断で Fail Over 動作を起こす必要があります。

仮想マシンのサービスレベルを守るための vSphere HA の設定

Virtual SAN 環境で vSphere HA を利用した場合、vSphere HA のハートビートネットワークは Virtual SAN ネットワークを利用します。Virtual SAN 有効時、無効時の vSphere HA の初期設定は以下の通りです。

	Virtual SAN 無効	Virtual SAN 有効
ハートビートネットワーク	管理ネットワーク	Virtual SAN ネットワーク
ハートビートデータストア	複数ホストからアクセス可能なデータストア	・Virtual SAN データストアは利用不可 ・別途追加は勿論可能
隔離アドレス	管理ネットワークのゲートウェイ	管理ネットワークのゲートウェイ
ホスト隔離時の動作	パワーオンのまま	パワーオンのまま

vSphere HA には以下の通り 4 つの状態があり、vSphere HA による仮想マシンの Fail Over は、ホスト障害の際、もしくはホストが隔離状態になった場合に発生します。

・正常な状態

・ネットワークパーティションの状態：

互いのハートビートのみが途切れた状態。ハートビートデータストアによりお互いの稼働が確認され、かつ、両者共に隔離アドレスへの Ping も成功する場合はこの状態となります。クラスタは分割され、それぞれが vSphere HA クラスタとして機能します。仮想マシンの Fail Over は起こりません。Virtual SAN 構成の場合、ハートビートデータストアが無い構成も想定されますが、その場合は、vSphere HA レベルのスプリットブレインとなり、仮想マシンの 2 重起動が起こる可能性があります。

・ホスト隔離の状態：

互いのハートビートが途切れ、かつ、障害ホストの隔離アドレスへの Ping が失敗する場合この状態となります。仮想マシンの Fail Over はオプション設定により可能（初期設定は無効）です。

・ホスト障害：

仮想マシンは Fail Over します。

ネットワーク障害の際には、ネットワークパーティションではなく、ホスト隔離の状態にすること、及び、隔離時に仮想マシンの Fail Over が起こるよう設定を施しておくことが必要となります。このため、以下の 2 点の設定変更を実施します。

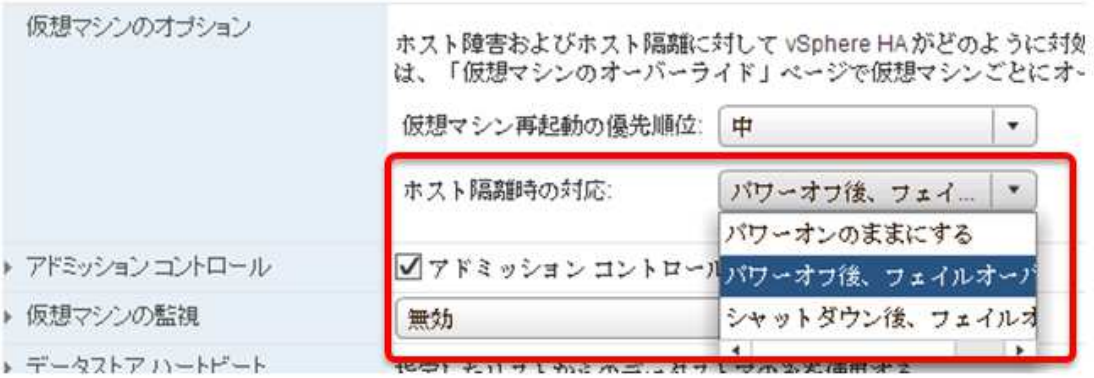
1.vShere HA をネットワークパーティションではなく、ホスト隔離状態にする

Virtual SAN ネットワーク障害時に、隔離アドレスへの Ping 応答を切断する必要があります。このため、vSphere HA の詳細オプションを利用して隔離アドレスを Virtual SAN ネットワーク側に変更します。

das.usedefaultisolationaddress = false

das.isolationaddressX = <Virtual SAN Network より確実にアクセス可能な IP>

2.vSphereHA のオプション設定で、ホスト隔離時の対応を、”パワーオフ後、フェイルオーバー”に設定する



この 2 つの設定を施しておけば、Virtual SAN ネットワークが切断された際、vSphere HA が隔離状態となり、隔離されたホスト上の仮想マシンが、正常なホスト上に Fail Over され、仮想マシンサービスが復旧します。
ハートビートデータストアに関しては必須ではありませんが、もし存在すれば、相手の状態を正確に把握することが可能となるため、仮想マシンの 2 重起動を防止することが可能となります。

設定変更の影響

上記の設定は必ずしも全ての仮想マシンの稼働の最大化を図る物ではなく、仮想マシンに想定されるサービスレベルを守るための物です。ここで言う、仮想マシンのサービスレベルとは、

“許容障害数以内のホスト障害であればサービスが継続、もしくは、 vSphere HA の機能で仮想マシンが Fail Over してサービスが復旧すること”

を意味します。初期設定では、特にネットワーク障害の際にこの期待値に添えない（仮想マシンの Fail Over が起こらない）可能性がありますが、前記した設定を施す事により仮想マシンに想定されるサービスレベルを守ることが可能となります。

ここで 1 つ心得ておきたいことは、許容障害数を超えてしまった仮想マシンにとっては、必ずしもこの設定が可用性を向上する物ではないことです。例えば、前記した例で、許容障害数 = 0 で作成された青色の vmdk ファイルは、esxi-01 のネットワーク障害時でも稼働は可能ですが、上記設定に変更してしまうと仮想マシンはパワーオフしてしまいます。今回ご紹介した設定は、仮想マシンに想定されるサービスレベルを守るための設定と理解下さい。

まとめ

今回は、Virtual SAN の障害時の動き、及び、Virtual SAN 上の仮想マシンの可用性を設計者の意図した通りに担保するための設定についてご紹介させていただきました。繰り返しになりますが、Virtual SAN ネットワークはチーミングによる冗長化が可能ですし、その重要性を考えると冗長化は必須です。その冗長化でも対応できない万が一の際にも、仮想マシンのサービスレベルを担保するためのデザイン手法として本資料をご利用頂ければ幸いです。

参考情報

VMware Virtual SAN & vSphere HA Recommendations
http://blogs.vmware.com/smb/2014/10/vmware-virtual-san-vsphere-ha-recommendations.html