

事業継続のための SAN エクステンション設計

2009 年 08 月



世界的にも、地震や風水害、津波などの自然災害、システム障害、セキュリティ被害、暴動、テロ、さらには新型インフルエンザのパンデミックなど企業を取り巻きリスクは増加する一方です。最近では、企業のグローバル化も手伝って、これらのリスクは見過ごせないほど大きなものになっています。

ビジネスにおける情報システムへの依存度が増大している今、IT 部門においても情報システムの停止が事業の中断に直結するリスクとなっており、事業継続のための早急な対応が重要な課題となっています。

事業継続の観点から考えた場合、最も重要なものはデータであり、保持しているデータを予期せぬ事象による破壊から保護する必要があります。そのためのソリューションとしては、以下の 3 種類が挙げられます。

- (1) バックアップによるデータ保護: Tape や Disk に取得するバックアップ
- (2) ストレージ間でのデータ保護: ストレージ機能を用いたデータのミラーリング
- (3) データセンター間のデータ保護: データセンター間での同期/非同期データレプリケーション

その中でも、(3) のデータセンター間のデータ保護を検討する企業が多く見受けられます。遠隔地に BC/DR (Business Continuity: 事業継続、Disaster Recovery: 災害対策) 用データセンターを用意し、データセンター間でデータ保護を実施するという広域災害を想定したソリューションです。これらの具体的な手法としては、ファイル転送やミドルウェア機能、ストレージ機能などがありますが、極力短い停止時間 (目標復旧時間: RTO) と極力直近のデータまで復旧させる (目標復旧時点: RPO) ことを目標とした場合、ストレージ機能で実現するストレージ レプリケーションが適しています。

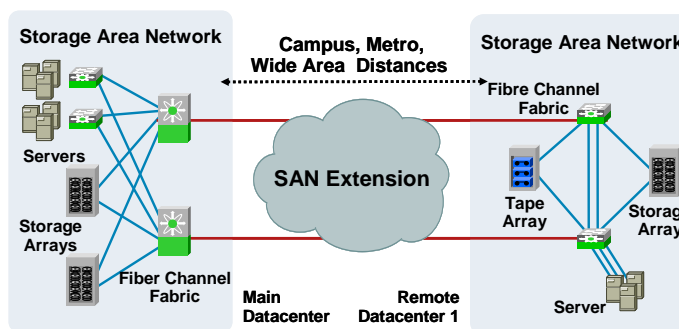
ストレージ レプリケーションは、広域災害を想定してストレージを遠隔地にも置き、ストレージの機能を用いてデータの同期または非同期のレプリケーションをするものです。また、遠隔地の BC/DR 用データセンターには、ストレージ以外にサーバやネットワークなどの機器をリカバリレベルや予算に応じて配備することになります。

では、事業継続を考える上で重要なソリューションであるストレージ レプリケーションを導入する際に、どのような観点で設計すべきなのでしょう。単純にレプリケーション機能のあるストレージ同士をネットワークで接続すれば良いのでしょうか。ネットワークで接続する際に何を考慮すれば良いのでしょうか。このホワイトペーパーでは、ストレージ レプリケーションの導入にあたり、ストレージ同士を接続する際ネットワークである SAN エクステンション (拡張 SAN) の設計ポイント、および設計時の考慮点をまじえながら技術的な解説をいたします。

ストレージ レプリケーションと SAN エクステンション

ストレージ レプリケーションとは、2台、または複数台のストレージ間でファイバ チャンネル インターフェイスを使用してデータをレプリケーションする機能です。この機能は、ストレージ装置内のソフトウェアとして提供されているため、同一ソフトウェアが動作するストレージ機種間でのみ実現することができます。

通常の SAN は、サーバからストレージを共有するために構築されるサーバとストレージ間のファイバ チャンネル ネットワークのことを言います。SAN エクステンション(拡張 SAN)とは、通常の SAN を拡張したもので、2つのデータセンター間のファイバ チャンネル ネットワークのことです。すなわち、ストレージ レプリケーション時の遠隔地にあるストレージ同士を接続するファイバ チャンネル ネットワークも、SAN エクステンションの1つです。

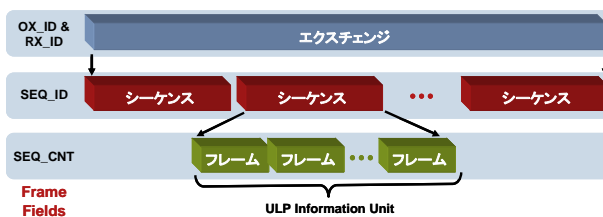


ストレージ レプリケーションのネットワーク プロトコル

- ファイバ チャンネル (FC:Fibre Channel)

サーバとストレージ間のストレージ ネットワークに使用されているファイバ チャンネル プロトコルをストレージ レプリケーションに使用します。このときのレプリケーション用のストレージ コントロールユニット(ストレージ コントローラ)は、ファイバ チャンネルをサポートしている必要があります。

ファイバ チャンネルにおけるデータ伝送構造を説明します。フレームとは、リンクに送られるデータの最小通信単位です。シーケンスは、1つの操作に関連する1つまたは複数のフレームで構成されています。エクステンジは、1つの操作に関連する、1つ以上の複数シーケンスから構成されます。たとえば、データを読むための SCSI コマンド、読み出したデータの転送、SCSI コマンドの完了ステータスは、3つの別々のシーケンスで、1つのエクステンジを形成します。



- FCIP (Fibre Channel over IP)

遠隔地に存在する2つの SAN アイランド同士を TCP/IP ネットワークを使用して相互接続するためのトンネリング プロトコルです。接続形態としては、基本的には、ポイントツーポイントで接続されます。トンネリング技術を使用するため、TCP/IP ネットワークではファイバ チャンネルのフレームをカプセル化することにより、ファイバ チャンネルを意識しません。データセンター間のネットワークを IP ネットワーク サービスで構成する場合、FCIPをサポートしているファイバ チャンネル スイッチにてファイバ チャンネル プロトコルを FCIP プロトコルに変換して転送します。



SAN エクステンション（拡張 SAN）の構成要素

一般に普及しているストレージ レプリケーションにおける SAN エクステンション(拡張 SAN)の構成要素は、ストレージ間のデータ伝送用通信機器、各種ケーブル、通信キャリアの提供するネットワーク サービス（広域ネットワーク、または光ネットワーク）、各種機器の保守運用のための監視ツール、および監視用ネットワークで構成されます。

- **データ伝送用通信機器**

ストレージ間のデータ伝送用通信機器としては、ファイバ チャネル スイッチ、または DWDM があります。

- (1) **ファイバ チャネル スイッチ**

ファイバ チャネル スイッチは、ストレージからのデータを IP ネットワーク サービスで転送する際に使用します。サーバとストレージ間のネットワークである SAN と同様のスイッチであり、複数のストレージ ポートの集約や、広域ネットワークに対応するための FC → IP へのプロトコル変換、およびデータの効率的転送のための圧縮などを行います。

シスコの MDS は、IP ネットワークを使用してファイバ チャネルを透過的に通すための FCIP (Fibre Channel over IP) を提供します。すなわち、ファイバ チャネルと IP ネットワークのポートをもっており、ゲートウェイとして機能します。



Cisco MDS9500シリーズ

- (2) **DWDM (Dense Wavelength Division Multiplexing)**

ストレージからのデータをファイバ チャネル プロトコルのまま、光ファイバ ネットワーク サービスで転送する際に使用します。DWDM は、既存の光ファイバ ネットワーク網を使用して伝送量を増大させるために利用されています。高密度波長分割多重方式により光ファイバ ケーブル上で波長の異なる複数の光信号を同時に伝送し、多重化することで広帯域化を実現します。



Cisco ONS15454 MSTPシリーズ

- **運用のための監視ツールと監視ネットワーク**

SAN エクステンションを構成する各種機器を監視するためのツールと監視用ネットワークが必要です。各機器は通常 SNMP や HTTP で管理可能なツールを提供しています。

監視方法として、機器個別に監視するか、集中的に監視するなどを決め、それぞれの監視ツールの監視項目で十分であるかも検討します。これらは、直接業務に影響が出ることはありませんが、事業継続を支える重要な構成要素であるため、通常のネットワーク機器と同様に監視する必要があります。

また、通常時から両データセンター共に監視するか、アクティブなデータセンターのみ監視するかなどの監視運用も決める必要があります。

ストレージ レプリケーション方式とあわせて検討すべき項目

ストレージ レプリケーション ソリューションを検討する際には、ストレージ レプリケーションの方式とあわせて検討すべき項目があります。以下にそれぞれの検討項目に関して説明します。

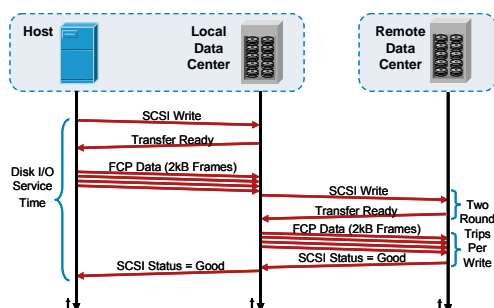
• ストレージ レプリケーション方式

まず、ストレージ レプリケーション方式を決定する必要があります。ここでは、一般に使用されているストレージ レプリケーション方式を紹介します。

(1) 同期レプリケーション方式

ホストからの書き込みデータは、プライマリのストレージに書き込まれ、ストレージ機能を使用してセカンダリのストレージに転送されます。そして、セカンダリは正常に書き込まれたことをプライマリに通知し、プライマリはホストに対して終了ステータスを送信します。すなわち、プライマリとセカンダリのストレージのデータは基本的に全く同じとなります。したがってリカバリ時に使用するデータにはズレがありません。

その反面、遠隔地にあるセカンダリのストレージまで書き込みに行くため、書き込み I/O ごとにパフォーマンス低下という影響を与える可能性があります。

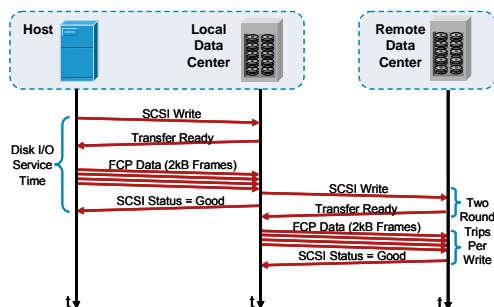


(2) 非同期レプリケーション方式

ホストからの書き込みデータは、通常通りプライマリのストレージに書き込まれます。そして、セカンダリのストレージに書き込むための転送用キューに入れられ、タイミングを見てトラック単位で転送されます。転送用のキューにたまっているデータが少なければ、セカンダリに即時に反映されることになります。パフォーマンスへの影響が少ないことがメリットですが、セカンダリへのデータ反映タイミングが遅れるため、災害時のタイミングによりデータロスが起こる可能性があります。

また、転送される順番と I/O の書き込み順番が保証されていないため、1 つのデータが異なるトラック上に格納された場合などではデータの整合性が保障できない可能性もあります。

そこで、上記のデメリットを改善するために、最近では一定時間内に発生した整合性を取った書き込みをキャッシュにキャプチャし、まとめてセカンダリのストレージに転送する方式も出ています。この方式では、パフォーマンスを犠牲にすることなくデータ消失の危険性を最小限に抑えることが可能なソリューションとなっています。



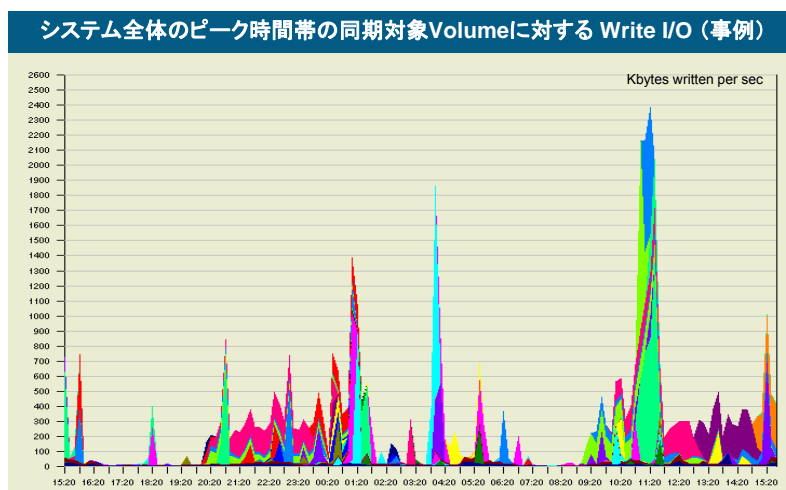
レプリケーション方式と回線帯域の見積もり

次にレプリケーションに必要な回線帯域を見積もります。見積もる際は、日次や月次などの運用サイクルや時間帯により書き込み I/O 数が大幅に異なることも多いため、運用に則した計画的な調査が必要となります。また、データ伝送用通信機器のデータの圧縮機能の圧縮効率により、見積もった帯域よりさらに少ない帯域で済む場合もあります。実際のデータを用いて圧縮効果を確認することをお勧めします。

(1) 同期レプリケーション方式

同期レプリケーション方式の場合、転送するデータの単位はホストからの書き込み I/O、すなわちブロック単位となります。一般的なデータベースのブロック長は、4K バイトや 8K バイトなどです。

同期レプリケーションで使用する回線帯域を見積もるには、レプリケーション対象システムのピーク時間帯の書き込み I/O 数 × ブロック長 = 書き込み I/O 帯域として積み上げることにより、必要帯域を算出します。

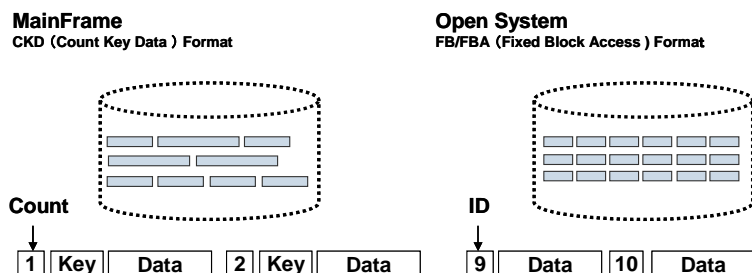


(2) 非同期レプリケーション方式

非同期レプリケーション方式の場合、ストレージがレプリケーション用に転送するデータの単位はディスクのトラック単位となります。一般的なディスクフォーマット方式の 1 トラックは、FBA 方式で 32K バイト、CKD 方式で 56K バイトです。

非同期レプリケーションでは、ストレージの負荷状況にあわせて完全な非同期タイミングで転送するものと、一定時間おきに更新されたトラックのみをキャプチャして転送するものなどがあります。

非同期レプリケーションで使用する回線帯域を見積もるには、同期レプリケーションと同様にレプリケーション対象システムのピーク時間帯の書き込み I/O 回数 (または変更トラック数) × トラック長 = 転送帯域として必要帯域を算出します。ただし、書き込み I/O の場合には同一トラックへの I/O も積み上げられてしまうことに留意が必要です。



● レプリケーション方式とネットワーク サービスの選定

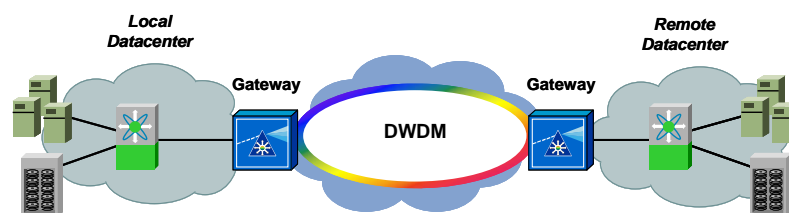
ストレージ レプリケーション方式と必要な回線帯域が決まったら、次にストレージ レプリケーション方式にあわせたネットワーク サービスの選定を行います。

(1) 同期レプリケーション方式

同期レプリケーション方式の場合、書き込み I/O が遠隔地にあるセカンダリのストレージまで転送されるため、業務アプリケーションのパフォーマンス低下となる可能性があります。

そこで、パフォーマンスの悪化を最小限にとどめるためには、通信キャリアから提供される回線サービスを高帯域で低遅延の光ファイバネットワークにすることを勧めます。

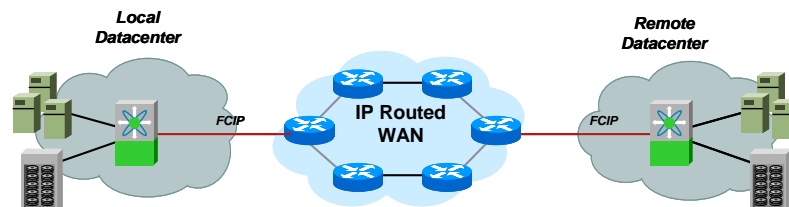
また、伝送距離が長くなるほど伝送遅延も大きくなるため、BC/DR 用データセンターの場所を決定する際の 1 つの要件として、距離と伝送遅延の関係も考慮に入れる必要があります。距離と伝送遅延に関しては、「同期レプリケーション時の距離による I/O 遅延」の項で説明します。



(2) 非同期レプリケーション方式

非同期レプリケーション方式の場合、ホストからの書き込みデータは、通常通りプライマリのストレージに書き込まれるためパフォーマンスへの影響が少ないことがメリットです。

完全に非同期なタイミングまたは一定時間間隔でまとまったデータを転送するため、同期レプリケーションほど遅延を気にする必要がありません。そのため、IP ネットワーク サービスで必要十分な帯域を使用することが多いのが現状です。ただし、IP ネットワーク サービスについては、多くのサービスが通信キャリアから提供されていますので、適切なサービスを選択することが重要となります。



SAN エクステンション設計時に考慮すべき項目

ストレージ同士を接続するネットワークである SAN エクステンション (拡張 SAN) を設計する際に考慮すべきその他の考慮点について、以下に示します。

同期レプリケーション時の距離による I/O 遅延

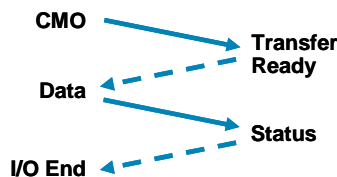
同期レプリケーションの場合の書き込みは、以前はローカルのストレージに書き込んで終了しましたが、セカンダリのストレージまで書き込みに行ってその完了をもって終了します。

すなわち、書き込み I/O はストレージ間の距離による伝送遅延が発生することになります。

距離による伝送遅延時間を算出する場合は、ファイバケーブル内での光の速度に基づいて計算します。1 km につき 5 μ sec、すなわち 1 ms で 200 km 進むことになります。ストレージ間の距離を 100 km と仮定した場合、片道 0.5 ms の遅延、往復の RTT (Round Trip Time) では、1 ms となります。

さらに、ホストやサーバからの書き込み I/O シーケンスは、ファイバ チャンネルのため 2 RTT の時間が必要となりますので、100 km 離れたデータセンター間の距離による遅延時間は 2 ms になると想定されます。

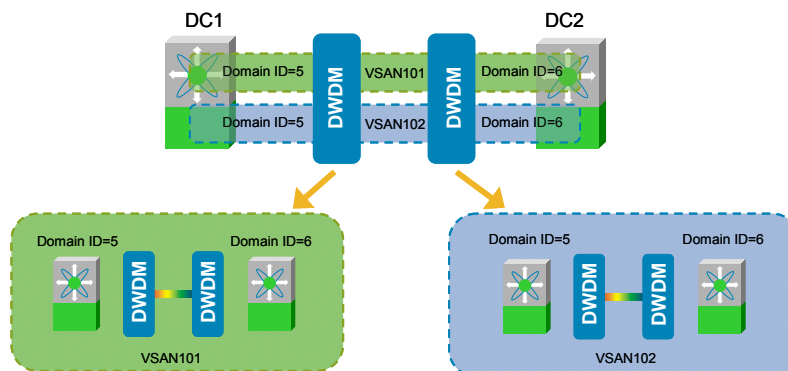
そのため、同期レプリケーションの場合には、この伝送遅延がトランザクションやバッチジョブにどれだけの影響を与えるか、事前に概算予測を実施する必要があります。



ファブリックの分割

SAN エクステンションはファイバ チャンネル ネットワークのため、デフォルトでは 1 つのファブリックとなります。1 つのファブリックには、1 つのファブリック サービスが提供されます。ファブリックへの接続状態やアクセス制御情報 (Zoning) に変更があった場合には、RSCN (状態変更通知サービス) が発行されます。この RSCN はブロードキャストで配信されるため、同一ファブリック内のすべての機器が影響を受け、ファブリックの再構築が行われます。

この影響を回避する方法として、シスコでは VSAN という機能を提供しており、お互いに影響させたくないグループごとに VSAN を定義し、ブロードキャストを分割することが可能です。転送するデータの重要度や種別ごとにグルーピングしたり、回線の冗長構成別でグルーピングしたり、運用や構成にあわせて分割設計することをお勧めします。

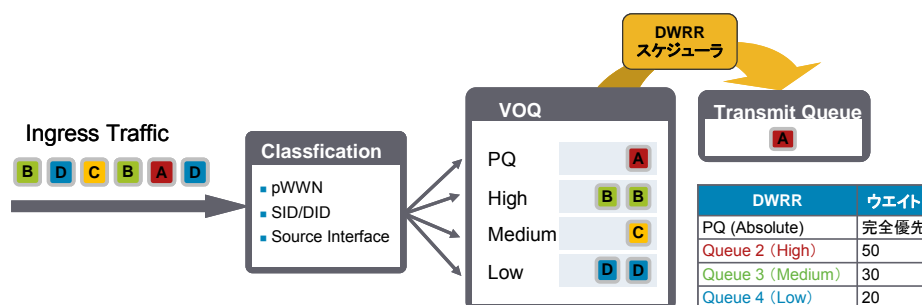


● 転送データの優先制御

レプリケーション方式ごとに必要な回線帯域の見積もりを実施しますが、この必要帯域の中でプライオリティを付けたい場合には、以下の方法が考えられます。

- (1) レプリケーション構成にて、優先制御が必要な回線を物理的に分割する
- (2) ファイバ チャンネル スイッチにて優先制御(QoS)の設定を実施する

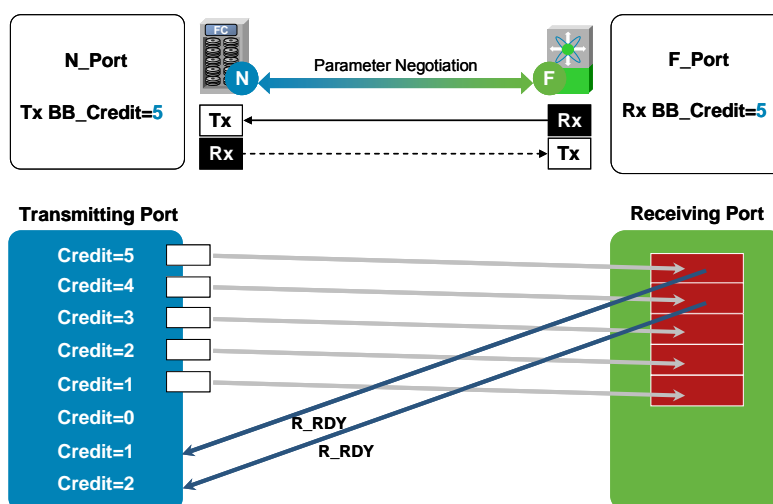
たとえば、同期レプリケーションと非同期レプリケーションの混在環境で上記の優先制御を行わない場合に、見積もった回線帯域を超えるデータが転送されたとします。アプリケーションのレスポンスの悪化や、最悪の場合タイムアウトとなり、エンドユーザにまで影響が及ぶ可能性があります。転送データの業務的なプライオリティやレプリケーション種別などをあわせて考慮して、転送データの優先制御を検討することをお勧めします。



● BB クレジット (Buffer to Buffer Credit)

BB クレジットとは、ファイバ チャンネルでの物理的に接続されたポート間でのフレームの受け渡し制御(フロー制御)を行うためのポートの受信バッファ数です。事業継続のためのレプリケーションをファイバ チャンネル プロトコルで転送する場合、長距離データ伝送となることが多く、この BB クレジットの値が伝送効率に大きく影響してきます。

一般的には、2G のファイバ チャンネルなら 1 km につき 1 BB クレジットで計算します。すなわち、2G のファイバ チャンネルで 100 km 先にデータ転送する場合には、100 BB クレジットが必要です。そのため、事前にファイバ チャンネル スイッチのデフォルト BB クレジット値を確認しておくことをお勧めします。



● セキュリティ

ベーシックなセキュリティ機能として、SAN エクステンション環境においても、ファイバ チャネル スイッチによるアクセス制御 (Zoning) があります。ストレージの WWN 同士で実施する WWN ゾーニングと、ファイバ チャネル スイッチのポート同士で実施するポート ゾーニングの 2 種類があり、運用方法やセキュリティの観点から、どちらの方式を採用するかを決定します。

シスコではセキュリティを強化する機能として、ファイバ チャネル スイッチのポートに対するログイン制御を行う Port Security や、ファイバ チャネル スイッチ間の接続制御を行う Fabric Binding などを提供しています。

● 各種タイムアウト値

SAN エクステンション内の各種タイムアウト値を確認し、どの機器で障害が発生しても障害検知と復旧が正しく行われるように整合性を確認する必要があります。

特に、同期レプリケーション時には、距離による書き込み I/O の遅延が発生するため、アプリケーションやデータベースのタイムアウト値まで考慮に入れなければなりません。

● 障害設計と障害検知

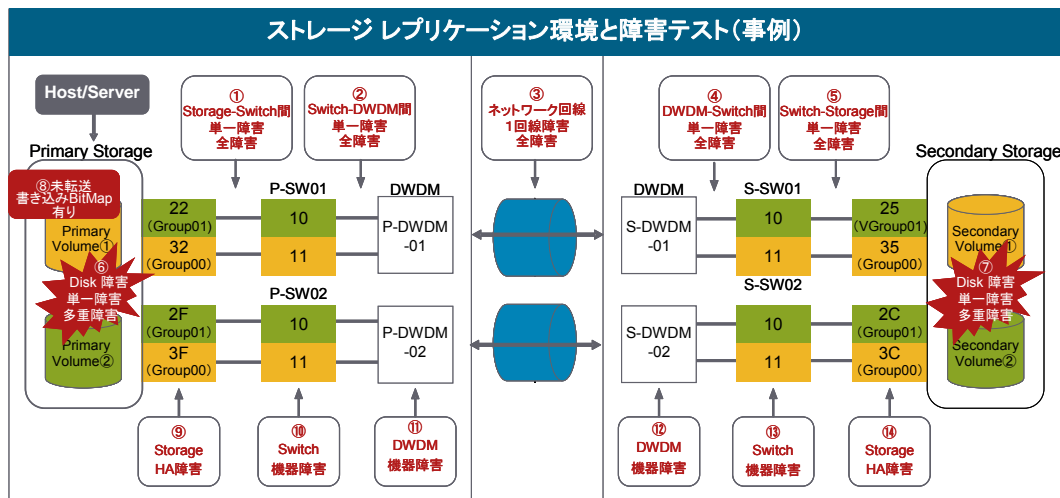
障害時設計においては、障害発生箇所とその影響を予測し、障害テストにて Timer の整合性を含めた確認を実施すべきです。また、既存の検知手段では運用要件を満たさない場合には、必要に応じた仕組みを構築することも検討します。

たとえば、あるストレージ メーカーの同期コピーでは、レプリケーション用のリンクが切断した場合、プライマリのストレージへの書き込み I/O を継続させるための機能としてプライマリのストレージ内に未転送の書き込みビットマップをもち、障害復旧時にはそのビットマップにしたがって未転送のトラックを転送することで差分同期を実施します。

このときのレプリケーションの運用要件として以下の事象を認識する必要がありました。

- (1) レプリケーション用のリンク障害により、縮退運用をしていること
- (2) 障害復旧後のプライマリとセカンダリのストレージ内容がまだ完全に一致していないこと

このような場合には、ストレージ ベンダーから提供されているものだけでは十分ではなく、別途チェック用プログラムなどの仕組みを構築することになります。



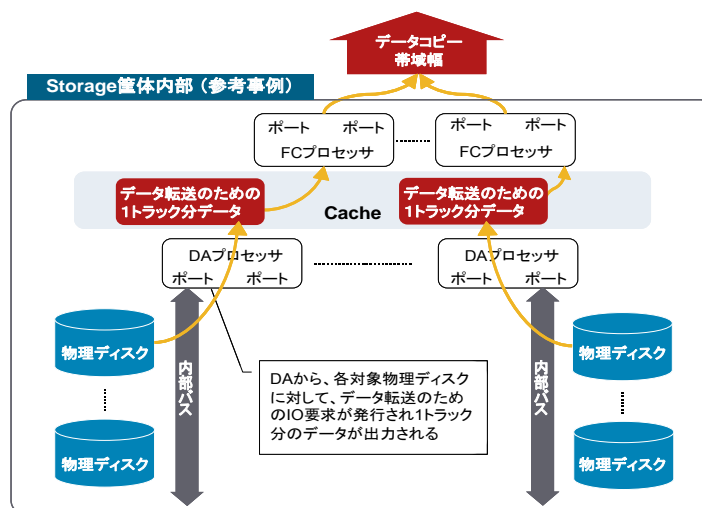
SAN エクステンション以外の考慮点

ストレージ レプリケーションを設計する際には、SAN エクステンション以外にも考慮すべきポイントがあります。代表的な考慮点を簡単に説明します。

・ 非同期レプリケーション時のストレージ内部の転送モデル

ストレージ レプリケーション方式のところで説明したとおり、非同期レプリケーションの転送単位はトラックとなります。また、ホストからの書き込み/読み込みデータとディスクから読み取った転送のためのデータが、ストレージ内部のバスやキャッシュ メモリ、およびディスク コントローラなどを共有します。そのため、レプリケーションを行う前よりストレージ内部の負荷状況が高くなることが想定されます。負荷状況が高くなるということは、通常のストレージに対するパフォーマンスが悪化する恐れがあります。

非同期レプリケーションを実施する際には、ストレージ内部の転送モデルを確認し、レプリケーションを実施したときの負荷を想定しておくことをお勧めします。

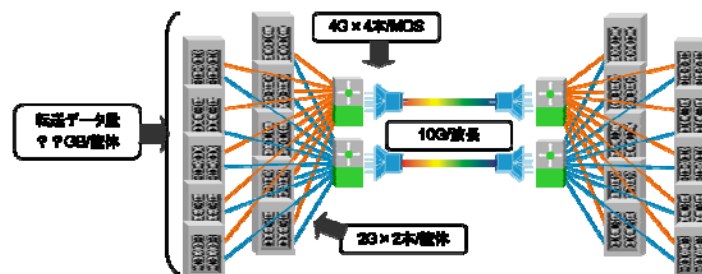


・ 初回レプリケーション時の影響

同期でも非同期レプリケーションでも、すべてのストレージにおいて最初の1回目は非同期モードでプライマリのボリュームとセカンダリのボリュームの同期が実施されます。初回レプリケーションを対象ストレージ一斉に開始した場合、すべてのストレージのすべてのボリュームが対象となるため、転送すべきデータ量が膨大となり、ボトルネックが発生します。そのボトルネックと初回レプリケーションにかかる時間を明確にし、対応策を検討する必要があります。

また、非同期レプリケーション時のストレージ内部の転送モデルで確認したように、ストレージ内部の負荷が高くなった場合には、初回レプリケーション時に本番環境へも影響を及ぼす恐れがあるため、何らかの対策を講じなければなりません。

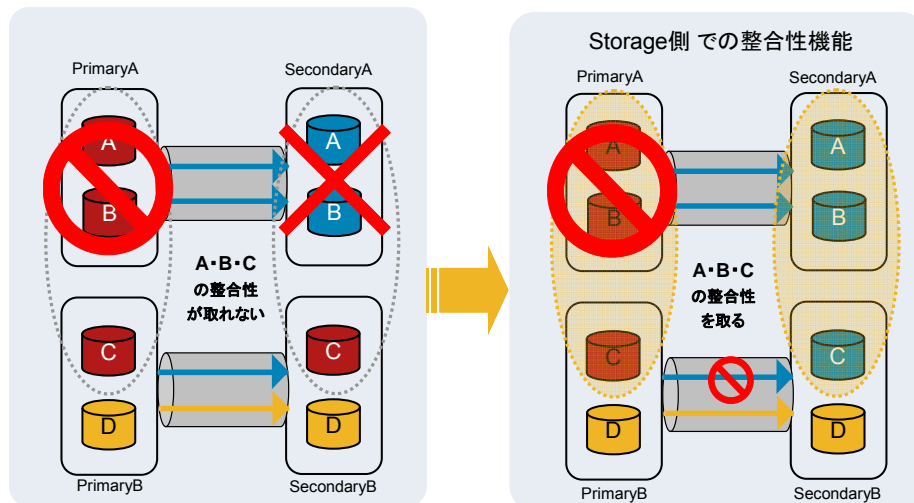
- ・ 初回コピーを複数回に分けて実施する(ボリューム数や筐体数を制限)
- ・ ストレージ内部での調整が可能であれば実施する、など



● ストレージをまたがるボリューム間の整合性

データベースなどでは複数のストレージをまたいでデータが格納されるため、ストレージの障害時に一部の領域だけが更新され一部が更新されないような事象が発生し、データの整合性が取れなくなります。このようなストレージをまたいで整合性を取るべきデータが存在する場合のレプリケーション設計では、プライマリ サイトでのストレージ障害時でもセカンダリ サイト内で整合性が取れるように考慮すべきです。

レプリケーション機能において、ストレージの筐体障害時に Volume 間の整合性を取る仕組み、たとえば整合性が必要な Volume 同士をグループ化しておき、どこかで障害が発生した場合すべての更新を実施しない、などの機能があればストレージ側にて実施します。



ストレージ側に該当する機能がない場合は、代替手段を検討しなければなりません。

シスコのファイバ チャンネル スイッチでは、トラッキング対象ポートを監視して、リンク状態の変化を検出した場合にリンク対象ポートをダウンさせる Port Tracking という機能を提供します。



- 1) tracked port = fcip1がダウン
- 2) linked port = fc3/1をMDSがダウンさせる
 - indirectリンクの障害に即対応、ストレージ側に通知する機能
 - ストレージの機能でindirectリンク障害に対応するより早いリカバリ動作が可能

● バックアップ取得

事業継続用のストレージ レプリケーション環境を構築したとしても、バックアップの取得は必要です。災害時などの物理的破損の場合は、すべてのデータを直近の状態に戻す必要がありますが、オペレーションミスなどの論理的破損の場合は、一部のデータがある時点の状態に戻す必要があります。

事業継続のためのストレージ レプリケーションの環境では、物理的破損に対してのみ効果を発揮します。プライマリのストレージに書かれたデータはセカンダリのストレージに瞬時、または一定時間後に反映されますので、常に最新状態となっています。そのため、論理的破損のためには、運用にあわせたタイミングで戻せるようなバックアップを別途取得すべきです。

まとめ

本書では、事業継続のためのストレージ レプリケーション導入にあたり、ストレージ同士を接続するネットワークである SAN エクステンションを設計する際の考慮すべき技術的な解説を行いました。

ストレージから SAN エクステンションに接続するネットワークのプロトコルは、ファイバ チャネルが標準となっているため、ファイバ チャネル プロトコル特有の考慮点があります。特に、障害発生時の影響を綿密に考慮した設計を行う必要があり、障害発生時の迅速な検知方法と、障害に対する対処方法も事前に想定しておくべきです。最悪の場合、不測の事態が発生して事業継続が発動された際、事業継続用に用意していたデータや環境が障害により役に立たないということも考えられます。

また、ストレージのレプリケーション方式を、同期レプリケーション、非同期レプリケーション、同期と非同期の混合レプリケーションのどの方式にするのかにより、現行業務のパフォーマンスと RPO(目標復旧時点)に大きく影響します。特に、コストを抑えるために重要業務は同期、その他は非同期という混合レプリケーションを検討した場合は、ネットワーク回線帯域の見積もりや転送データの優先制御など、構成により検討が複雑となる場合があります。

ストレージ レプリケーションを導入する際は、RTO(目標復旧時間)や RPO(目標復旧時点)などの事業継続の要件と導入コストだけでなく、既存の業務システムへの影響と導入後の運用まで考慮した設計をすべきです。

最後に、このホワイトペーパーが事業継続のためのストレージ レプリケーション導入の際の参考となれば幸いです。

シスコシステムズ合同会社
アドバンスドサービス
データセンターネットワーキングプラクティス

©2009 Cisco Systems, Inc. All rights reserved.

Cisco, Cisco Systems, および Cisco Systems ロゴは、Cisco Systems, Inc. またはその関連会社の米国およびその他の一定の国における登録商標または商標です。

本書類またはウェブサイトに掲載されているその他の商標はそれぞれの権利者の財産です。

「パートナー」または「partner」という用語の使用は Cisco と他社との間のパートナーシップ関係を意味するものではありません。(0704R)

この資料に記載された仕様は予告なく変更する場合があります。



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>
お問い合わせ先：シスコ コンタクトセンター
0120-092-255 (フリーコール、携帯電話：PHS含む)
電話受付時間：平日10:00～12:00、13:00～17:00
<http://www.cisco.com/jp/go/contactcenter/>

お問い合わせ先