解 説 5

変わり身の術を実現するリモート・ストレージ

ー ストレージ基盤で障害と災害に同時対応する遠隔クラスター技術 ー

昨今、大量のデータをいかに効率的に保管・管理するかがIT業界における大きな課題となっていますが、そのための一般的なアプローチは統合ストレージあるいはその先のプール化という、できるだけ集約して無駄をなくそうという考え方です。

集約するに当たってはデータの価値や性質を考慮する必要がありますが、情報・データは買い直すことができず、失った場合には企業の存続さえ脅かされる性質を持つ最重要資産ともいえるものが含まれていることに気付きます。そのような資産が集約して保管されるとなれば、可用性が極めて重要な要件となることはITの専門家でなくとも容易に想像できると思います。

本稿では、ストレージにおける可用性についてあらためて考えるとともに、物理的なセンター施設や設備に依存しない隔地間でのストレッチ・クラスターを構成することで先進的な可用性を提供する、IBM System Storage SANボリューム・コントローラー（以下、SVC）によるスプリットI/Oグループ・クラスタリング（以下、Split I/O Group Clustering）について解説し、その技術がIT環境にもたらすメリットや将来的な意味についても考察します。

❶ はじめに

深夜の仕事というと、かつては道路工事現場の作業員やコンビニエンスストアの従業員、あるいは海外金融市場で取引を行うディーラーなどを思い浮かべました。しかし、生活習慣の変化に伴うサービス競争やインターネットの普及、また最近では一般企業を含むグローバル化の進展や国際分業の促進などの要因が、24時間の稼動を必要とするシステムが当たり前という状況を作り出しています。実際のところ深夜にe-メールやオンライン・ゲームが使えなかったら、ほとんどの方が困惑するのではないでしょうか。それがたとえ無料サービスだったとしてもです。

このように、もはやシステムの連続稼働、高可用性（High Availability：以下、HA）は当たり前という世界になっているところへ、近年ではサーバーだけでなくストレージも使用効率向上のための統合が進展しています。そして統合ストレージとなることで、従来の個別に配置される構成と比べ、障害停止による影響範囲が広がり、メンテナンス停止の調整も大幅に困難となるなどの状況が生み出されています。

一方、最低限の災害対策（Disaster Recovery：以下、DR）である遠隔地へのデータ保管や、DRやデータ・バックアップ実施のための事業継続計画（Business Continuity Plan：以下、BCP）の観点からも、データは企業にとって失ってはならない重要な資産として位置付けられることはご存知の通りです［1］。しかし、昨今のデータ量増加の傾向と前記の連続稼働の要請から、BCPの基本中の基本であるバックアップを取ることさえ時間的に困難になりつつあります。

このような、データ量の増加に加え連続稼働を強く求められる環境において、特定のサーバー・プラットフォームやソフトウェアに依存せずにHAとDRを同時に実現しようという高度なチャレンジが始まっていますが、ストレージでの実装はサーバーとは異なる困難さがあります。ストレージはデータを蓄え、かつその保管データに対する整合性や一貫性を保証する必要があるということがその困難さの原因となっています。本稿ではその点について解説し、今日実現している解決策の例をご紹介します。

❷ ストレージの可用性とデータの整合性および一貫性保証

サーバーとストレージとの可用性に関する主な違いは、サーバーは最終的な保管データを持たないようにする構成が可能で、その場合使用するデータ（OSやミドルウェア、処理プログラムを含む）にさえアクセスできれば、アプリケーションは異なるサーバーを使って、また異なるセンターであっても、切り替えて使うことが可能であるという点にあります。実際にはさまざまな課題や考慮点があるのですが、基本的には同種のサーバーを用意すれば稼働でき

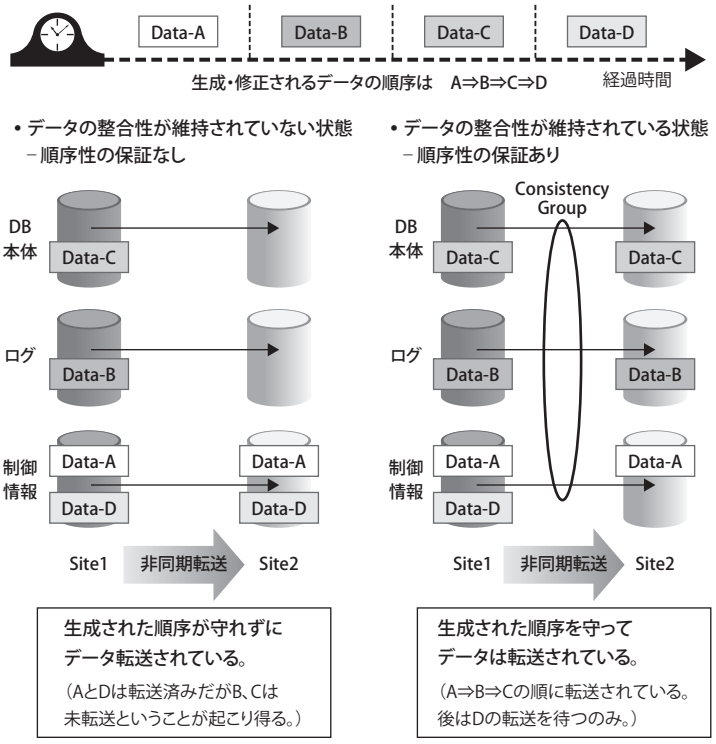


図 1. データの整合性

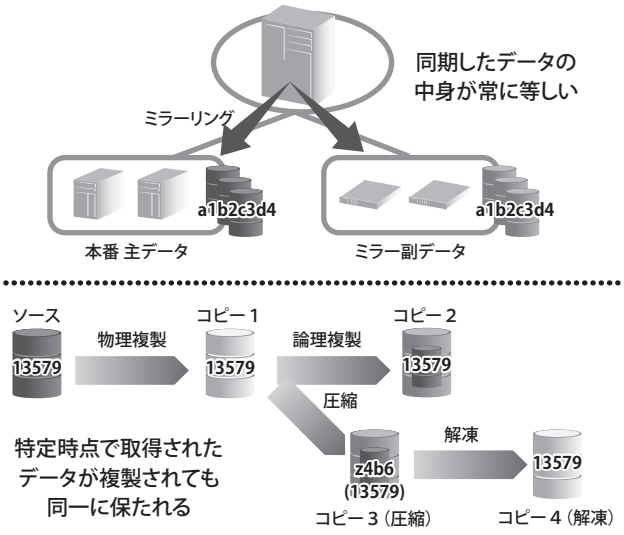


図 2. データの一貫性

そうだと理解できます。それに対してストレージの場合はどうでしょうか。ストレージは業務に必要なデータを保管する場所ですので、中身が空の同じ装置を用意するだけでは、業務で使用するデータが入っていないため、そのままでは使用できず、これでリカバリーしてもシステムを稼働できないことが容易に想像できます。

そもそもデータというものは発生時点ではユニークで唯一無二のものです。それがデータの重要度に応じて冗長ハードウェアに保管されたり、2 重書きやコピーが行われたりします。そして重要度が高いほど何重にも冗長化されるので、通常は多くのコピーを持つようになります。皆さんも大事な書類の複製やお気に入りの写真を焼き増しした経験があるのではないでしょうか。

またデータのもう 1 つの厄介な属性が、刻々と変化する（更新される）ものがあるということです。すべてのデータが一度作られると二度と変更されないものであれば扱いは比較的容易ですが、刻々と変化するという特性から、いつの情報か、その情報と関連する情報との整合性が保たれているのかということへの配慮が必要になります（**図 1**）。整合性の重要さは、例えば、カードで買い物をした場合に、月末の請求額が正しくても、買い物をした分のポイントが正しく反映されていないのは不都合であり、逆に請求額に誤りがあれば大問題となることなどから容易にお分かりいただけると思います。

この整合性の問題に加えて、可用性を高めるために更新の行われるデータに対して異なる装置間で複製（ミラーあるいはコピー）を行うとデータの一貫性（**図 2**）が損なわれる可能性があるということも問題になります。ミラーというと HDD を 2 つ用意して 2 重書きする RAID1 という技術を思い浮かべるかもしれませんが、ここで話題にしているのは装置内でのミラーではなく、制御装置の多重化や HDD を RAID 技術で保護したストレージ装置自体の停止までも想定した冗長性確保のためのミラー方法のことです。企業存続を維持するための大事なデータですから、確率が低い事象に対しても保険を掛ける必要があり、ストレージ装置を複数使った装置間複製を行うこともシステムによっては検討されています。

データの内容が変化している状態でこの装置間の複製を行うわけですから、その更新を複製物に対しても保証する必要があり、内容を同じに保つか差異の程度を管理するという、少し考えただけでもかなり大変な作業を行うことになります［2］［3］。

ストレージを物理的に複数用意してシステムとしての可用性を高める際にも、データの整合性および一貫性を保証する必要があることが、特に地理的に離れた場所で分散配置を行う場合の技術的困難さの原因となっています。

❸ SVCとは

世界各国の人々がそれぞれ異なる母国語で会話するように、ストレージは装置ごとに操作の画面やコマンドが異な

り、サーバーと接続するためのマルチパス・ドライバーも、装置ごとに専用のものを使用するのが当たり前となっています。それはさながら各国語に対応する専門の通訳をそれぞれ用意するようなものでした。最近ではOS標準のマルチパス・ドライバーを使用するケースも増え、統一できる環境が整いつつあるものの、現実にはインターオペラビリティー（装置間の相性）の問題があり、サポートされるバージョンが異なると同時接続がかなわないケースも生じます。また、同時接続できる場合でもファイバー・チャネル（FC）のアダプターあるいはポートを分ける必要があるケースが多く、とても自由に使えるという状況にはありません。

SVCはサーバーとストレージ装置の間に入ってサーバーから見たストレージ装置を仮想化、すなわち各装置の物理属性を隠ぺいすることで、共通のコマンドや機能を提供し、複数のディスク装置を使用する場合の厄介なインターオペラビリティーの問題を解決する、万能通訳者のような役割を果たします（**図3**）。

SVCの利用でコマンドや操作が統一されることにより、ディスク装置のコピー機能を利用したバックアップや災害対策のためのリモート・ミラーなどの仕組みの共通化、再利用が可能になり、新たなディスク装置を導入するたびに一から構築し直す必要がなくなります。また、複数種類のディスク装置が存在していても、サーバーからはストレージ装置としてはSVCしか見えませんので、サーバー側のFCのポートの共有や共通ドライバーの使用など、従来頭を悩ませていた問題が一気に解決します［4］。さらに、SVCはその稼働を止めずにハードウェアを追加して性能を向上するだけでなく、順に交換して最新のハードウェアに入れ替えたり、システムを止めることなくソフトウェアを更新して最新の機能を使用することまで可能にしています。このことは定期的な部分入れ替えを計画すれば、インフラの陳腐化を防ぐとともに従来は考えられなかったほど飛躍的にライフサイクルを伸ばすことが可能になることを意味しています。

SVCは物理的には**図4**のようにストレージのファイバー・チャネル・ネットワーク、いわゆるSAN（Storage Area Network）のスイッチを仲介にしてサーバーとストレージの間に入り仮想化を実現します。可用性の観点から制御装置を2台セットにして構成されますが、最大8台まで接続できる拡張性を持っています。この2台のセットをI/Oグループ、それを最大8台構成にする場合の全体をSVCクラスターと呼んでいます。

この各I/Oグループを構成する2台の制御装置のおのおのを、物理的に距離を離して配置することで、SVCクラスターを遠隔地で構成（ストレッチ・クラスター）し、拠

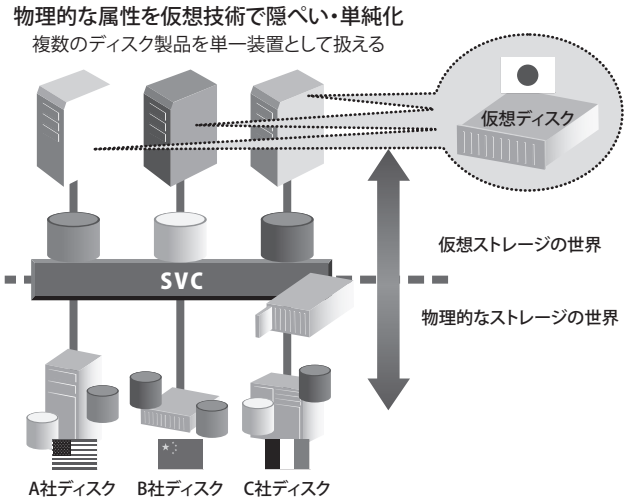
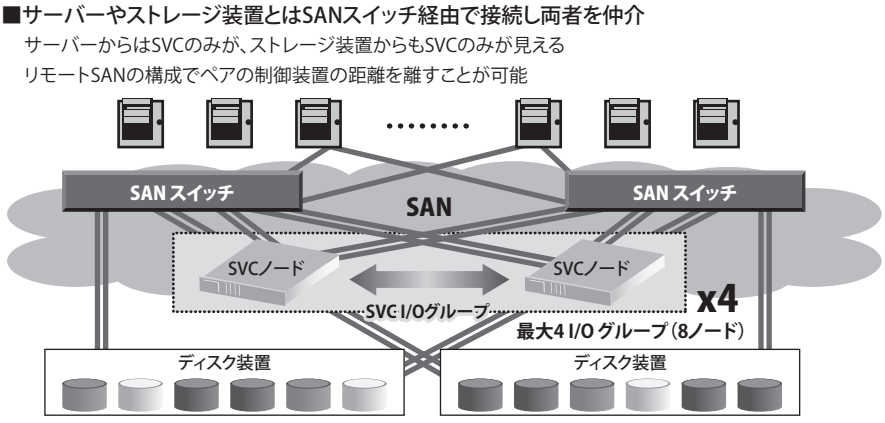


図3. SVCの概要



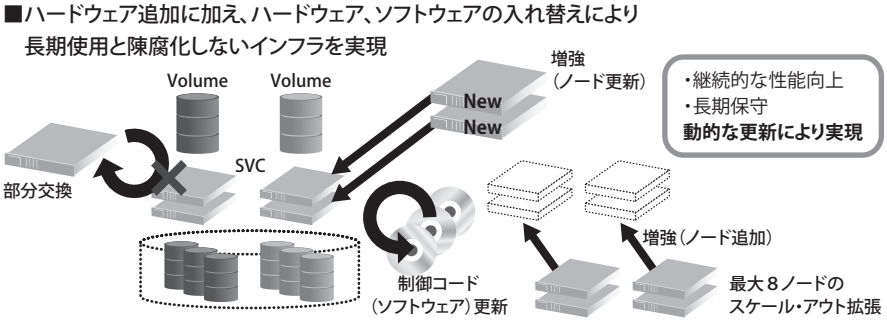


図4. SVCの結線と増強／保守イメージ

点間でのシステム可用性向上を果たすのが、今回ご紹介するSVCのSplit I/O Group Clusteringになります。

❹ 災対と高可用性を同時に実現するSVC Split I/O Group Clustering

HAやDRと聞くと皆さんは何を思い浮かべるでしょうか。立場によりさまざまだと思いますが、調達する立場であればどちらもリソースが余分に必要でお金が掛かりそうだというイメージを持つのではないでしょうか（**図5**）。また、DRシステムを検討したことがある方であれば、DRは切り替わる際にシステムが一旦止まってしまうということをご存知かもしれません。

そもそもDRを考えるとき、かつてはまず起こらないことであるという前提で検討されることが一般的であったため、対策の対象外となるシステムが多く、再開時間（Recovery Time Objective：以下、RTO）も数日や1週間などという条件がありました。しかしながら東日本大震災以降、災害は起こり得るという前提で考えられるケースが増え、要件が厳しくなったことにより、費用の増加と普段は使われないということがあらためてフォーカスされるようになっています。

同一センター内のHAであれば、正副両方のサーバーを稼働させて使用すること（Active-Active）も普通に行われていますが、DRに関しては距離の離れた2つのセンターの一方のシステムを本番とし、もう一方はスタンバイで停止させてデータだけを送っているという形態が多いのが現状です。また、DR用センターにはストレージだけ用意して業務用サーバーは置かないという構成も可能ではありますが、その場合は災害が起きてからサーバーを調達することになるためRTOが非常に長くなってしまうことやリハーサルが行えないなどの理由から、主要な業務用のサーバーをあらかじめ用意されるお客様が多くなっています。サーバー・リソースに関しては、災害が起こるまでの非活動の間は最小限のリソース契約で、実際に使う際にダイナミックに能力を拡張することができるようなソリューションも存在しますが、両方のセンターにサーバーとストレージがあって業務では一方しか使用していないという形態は、クラウド・コンピューティング、特にパブリック・クラウドというシステムの物理的な位置を意識しない構成が普及している現状ではいかにも無駄に思えてしまいます。

この無駄という感覚を少し掘り下げてみると、少し面白いことに気付きます。HAもDRもリソースが余分に必要になることは共通していますが、その余分なリソースが通常時に業務処理のために「使われるか」「使われないか」という点に関しては、リソースをハードウェアやソフトウェアなどの実体として見るか、能力という視点でとらえるかによって心象が異なっています。また、ストレージとサーバーでも、主にデータを蓄えるかどうかという違いによって差異があります。

筆者個人の主観も含みますが、どのレベルで無駄ととらえられるかを最近の経験から思い返すと次のようになります。まずサーバーのHAは、実体は使われる（Active-Active）形態が増え、可用性に価値を見いだすという考え方［5］も浸透しており、可用性が重要と思われるシステムでは当たり前のように使われています。そしてこのActive-Activeという形態によって無駄というイメージはそれほどクローズアップされていません。ストレージに関しても、制御装置部分はサーバー同様Active-Activeとなり、データ部分はRAIDなどで冗長化を確保する構造と

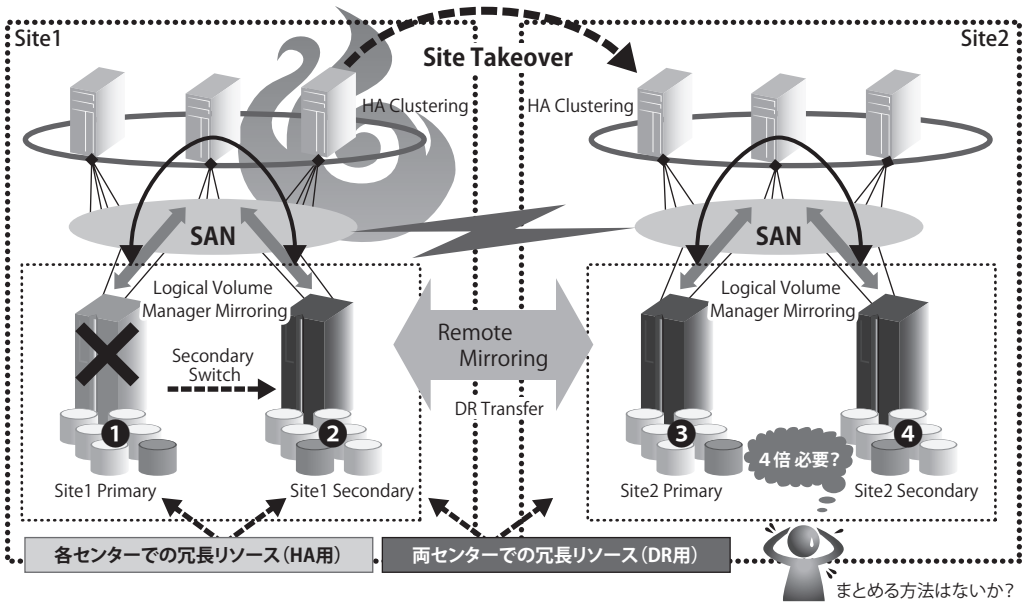


図5. 一般的なHAとDRの概念

なっており、ここまでは無駄というイメージにはつながっていません。しかし、多重化構造を持つディスク装置自体を複数用意して2重書きによる冗長性向上でのHAとなると、そこまで行うケースは非常に少なく、「そこまでする必要があるのか」という感覚あるいはお客様の言葉を借りると、このレベルまでの冗長性確保は無駄という感覚を持たれるケースが増えるようです。そもそもストレージの場合に冗長化したものを複数用意するという議論を行うのは、先に触れたように保管しているデータには失ったら再生できないという特性があるためです。そして、この特性のためにデータに対してはバックアップという考え方があり、失わないように保全が図られます。故に冗長化のコストに比べて、対象となるデータの価値が非常に高い場合とシステムが停止している時間による損失が非常に大きい場合のみ、ストレージ装置自体の複数用意が検討されます。

次にDRに関してですが、これは現在のソリューションではサーバーもストレージも被災しない限り必要とはならないという意味で余分なものであり、できれば費用を掛けたくないというのがどのお客様でも共通の感想です。

さらに、リソースを能力としてとらえた場合、縮退を許容するか否かでも違いがあります。HAもDRも事が起きた際に、それ以前と同じだけの処理をすべて継続できるようにと考えれば通常時には能力として余らせて持つ部分が大きくなり、それとは逆に、事が起きた際には一部の業務は継続しないあるいは処理の能力が低下することを許容するのであれば、余分な能力は最小限でよく、この場合は無駄と思える部分は少なくなります。

冗長性と縮退の考え方は航空旅客機や自動車のエンジンに例えると分かりやすくなるかと思います。旅客機は通常エンジンを複数持っており、そのうち1つのエンジンが停止しても巡航できるようになっています。エンジンが1つでも停止したら、海の上だろうと緊急着水するしかない旅客機にはとても怖くて乗れないと思うのが通常ではないでしょうか。つまりこの旅客機の例は、縮退はしないようにリソースを確保しているケースと同様に考えられます。これに対して、通常の自動車のエンジンは複数気筒ではありますがエンジンとしては1つしかありません。細部を議論するとさまざまな反論もあると思いますが、自動車の場合はエンジンが停止しても飛行機のように墜落するということはありませんので、エンジン単体の可用性は確保するものの最悪大きく縮退したり、停止した場合はあきらめるという割り切りの構造という見方ができると思います。このような違いはコストと影響範囲という視点も関係しています。

前置きが非常に長くなりましたが、HAとDRを一緒にして冗長化部分をマージするような構成を取るのがSVCのSplit I/O Group Clustering（**図6**）です。具体的にはHAのために2重化構成となっている各I/O GroupのSVC制御装置を、距離を離して異なるセンターにそれぞれ配置し、データ部分に関してはSVCのVolume Mirroring機能により両センターのデータが常に同じ状態となるように保ち、個別の障害だけでなくセンター全体の被災時にも制御装置部分とデータ部分を同時に切り替えることで、HAとともにDRを実現できる構成としています。このSplit I/O Group Clusteringによってストレージの部分のHAとDRが共通化されますが、一方がスタンバイあるいはセカンダリーといった業務非稼働状態で使われるのではなく、本番業務の観点でも両センター同時に稼働させることが可能となります。

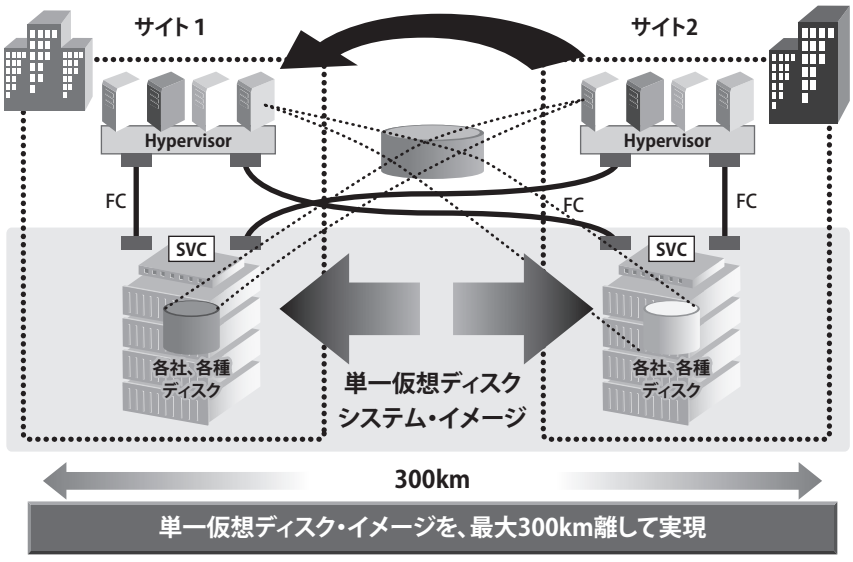


図6. SVC Split I/O Group Clusteringの構成イメージ（DRとHAを同時に実現）

もちろん従来の技術でも本番・DRの両センターで業務を稼働させ、災対用に保存の必要のあるデータをたすき掛けのように送り合って両センターで利用する形態はありましたが、HA部分はそれぞれのセンターで個別に実現しており、DR部分はDRが発動するとシステムの停止とリカバリーや再始動を必要とし、業務再開までには一定の時間を必要とする構成となっていました。これに対してSplit I/O Group ClusteringはDRが発動してもHA

同様に短時間の Wait のみでシステムを継続稼働させることをストレージのインフラ要件とするための技術です。

ただし、ここで「ストレージのインフラとして」と記載したのは、ストレージだけが切り替わる際にサーバーやアプリケーションが連続稼働するにはファイル・システム・キャッシュや DB バッファーといったサーバー側で管理されるリソースへの対応が併せて行われる必要があり［6］［7］、残念ながら現状では対応可能なサーバー環境は限られているからです。

複数のセンターにある機器をそのロケーションを意識せずに 1 つのシステムとして扱うという構想は以前から存在し、15 年程前にはその基礎となる技術が実用化されています。メインフレームの世界では 2 センターを同じ業務で相互利用するというプラットフォーム・レベル（あるいはインフラ・レベル）のソリューションが 1990 年代の終わりごろには存在していました［8］。

しかし、オープン系のプラットフォームでは、参照系の業務や業務の処理を分割したトランザクションを複数のセンターに送り込んで対応する形態のソリューションは存在しましたが、更新処理を行う業務を災害時に稼働したまま引き継ぐことはできませんでした。また、計画的に業務を稼働したまま異なったロケーションに移動する技術は、物理サーバー間で論理サーバーを移動する仮想化ソリューションの拡張という形で数年前には実用化されていますが、このソリューションでは距離の離れたセンター間での災害対策（非計画停止）には対応できていません。

このようなセンター間での HA と DR の同時実現が困難であることの背景には距離とその遅延の影響があります。現在、長距離接続のネットワークのバックボーン回線には物理的には光ファイバー・ケーブルが使用されていますので、その伝送速度の物理的な限界は光の速さであり、そのためどんなに高品質のネットワーク回線を契約しても理論的な速度の限界は光のスピードということになります。

ファイバー・ケーブル中の光の速度は真空中の光の速度約 30 万 km/ 秒の 70% ほどになりますので、100km 往復（200km）で 1ms 程度です。そのためデータのやり取りでの距離遅延を考える際は、100km 当たり 1ms が机上の RTT（Round Trip Time）概算値となります。しかし実際には、回線距離は直線距離より長くなることや 2 拠点の間に入る通信機器や中継機器での遅延もあり、一般に直線距離で約 600km となる東京 - 大阪間で遅延の簡易測定を行うと RTT は 10 ～ 20ms 程度となることが多いようです。これを HA 構成への影響として考えると、例えばサーバーおよびストレージをそれぞれ 2 台構成にして冗長化している場合に、サーバーとストレージ各 1 台を異なるセンターに配置すると、データ冗長化のために両センターのストレージ装置のデータを同じに保つには、書き込み処理があるたびに離れたセンターへデータを送ってその返事を待つ必要があります。このことにより処理性能が遅延の影響を受けて低下することになります。

実際には、遅延が 1ms の環境と 10ms の環境では返事を待って送る場合の転送効率が約 10 倍違うことになり、遅延が 10ms ある環境では、返事を待って行う処理は 1 秒間（1,000ms）に 100 回以上は行えないことにな

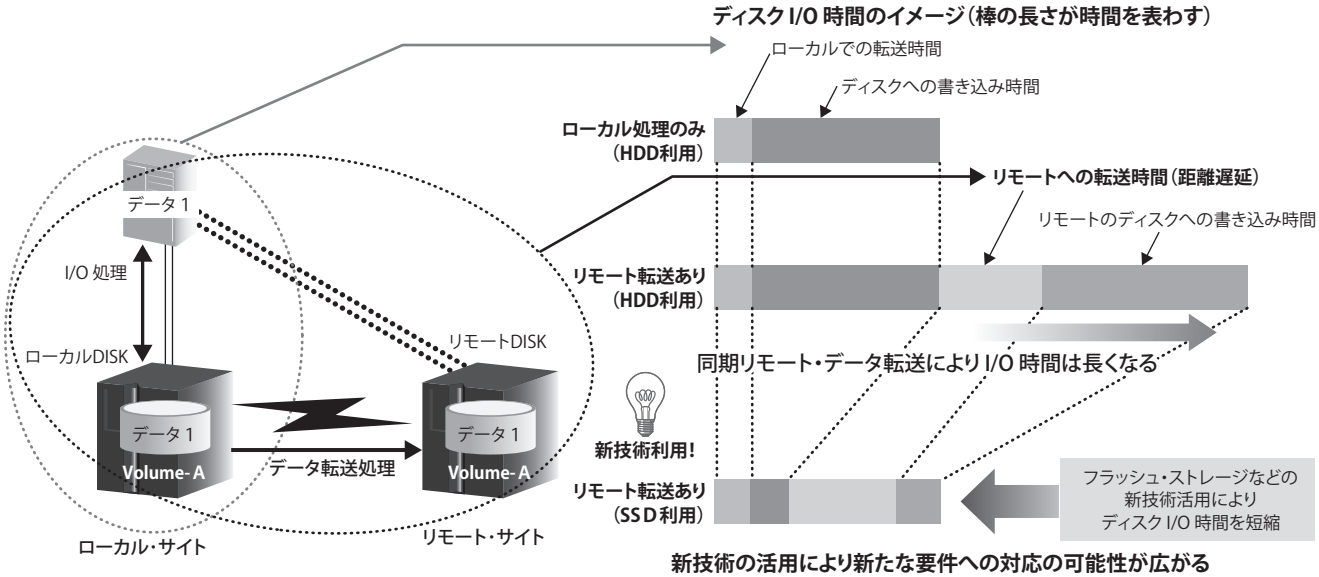


図 7. 距離遅延の影響を新技術活用により相殺

ります。昨今のサーバー、ストレージ環境では1秒当たり数千回から数万回といった読み書き処理（IOPS：I/O per Second）が行われているという現状とは2けた以上のギャップがあり、並列処理を考慮したとしてもその困難さを想像いただけるかと思います。

このためSVCのSplit I/O Group Clusteringでは、ストレージ間転送部分の返事を待つ処理をできるだけまとめて送って回数を減らしたり、1回の処理でのやり取りの数を減らしたりすることで、300kmまで距離を離して冗長化制御装置を配置できるようになっています。今まで困難と考えられていたことが具体的な製品機能として実現したということは、ロケーションを意識しないインフラ実現のための大きな第一歩といえるでしょう。

❺ まとめ

以前はお客様ごとの要件を反映した特注ともいえるインフラを用意するという手法が一般的でしたが、ハードウェアの性能向上と仮想化やクラウド・コンピューティングの進展に伴い、標準インフラを使用してシステムを統合することに加えて、システムの物理的な場所がどこにあっても構わないという考えが急速に広まっています。もちろん、多くのお客様が自社のデータや重要業務は自社所有のシステムでと考えられていることにはまだ変わりはありませんが、パブリックを含むクラウド・コンピューティングの利用可能性が検討されたことで、システムの棚卸しが行われ重要度や要件が整理されたお客様が増え、そのようなお客様では結果的にリモート・システム環境への適用度やインフラに掛けられるコストが可視化され、技術的な課題以上に制約となっていた高度なDRやHAの採用判断のための障壁が取り払われつつあります。

また、昨今大手企業のみならず、企業の買収、合併などの影響で複数システム、複数センターを運用するケースが増え、単なる統廃合だけではない効率的な複数センターあるいは複数システム運用に対するニーズが高まっています。こうした需要に応えるためにもSplit I/O Group Clusteringが有効となる可能性を持っています。

SSD（Solid State Drive）をはじめとするフラッシュ・ストレージを活用することで、多くの環境においては同じ接続条件であれば現状と同じ負荷を掛けながら数msのレスポンス・タイム短縮が可能となります。つまり数百キロ離れたセンター間でSplit I/O Group Clusteringを組むことにより受ける書き込みに対する数msの影響を、普及期を迎えているフラッシュ・ストレージのような新しいテクノロジーとセットにして相殺することが可能（**図7**）であり、現在はリスクを減らしてSplit I/O Group Clusteringのような新たな取り組みを行える絶好のタイミングともいえるのです。

[参考文献]

[1] Buecker, A., McConomy, D. Ferreira A. and Vora, N.: High Availability and Disaster Recovery Configurations for IBM SmartCloud Control Desk and IBM Maximo Products, IBM Redbooks ,http://www.redbooks.ibm.com/abstracts/sg248109.html

[2] Brooks, C., Leung, C., Mirza, A., et al.: IBM System Storage Business Continuity:Part 1 Planning Guide, IBM Redbooks, http://www.redbooks.ibm.com/abstracts/sg246547.html

[3] Brooks, C., Leung, C., Mirza, A., et al.: IBM System Storage Business Continuity:Part2 Solutions Guide, IBM Redbooks, http://www.redbooks.ibm.com/abstracts/sg246548.html

[4] Lovelace, M., Gebuhr, K., Gomilsek, I., et al.: IBM System Storage SAN Volume Controller Best Practicesand Performance Guidelines, IBM Redbooks, http://www.redbooks.ibm.com/abstracts/sg247521.html

[5] Marcus, E., Stern, H.: Blueprints for High Availability, John Wiley & Sons (2003).

[6] Bartkowski, S., Buitlear, C., Kalicki, A., et al.: HighAvailability and Disaster Recovery Options forDB2 for Linux, UNIX, and Windows, IBM Redbooks, http://www.redbooks.ibm.com/abstracts/sg247363.html

[7] Brumer J., Cecil R., Fabbri F. and Rees S.: Configuring geographically dispersed DB2 pureScale clusters, https://www.ibm.com/developerworks/data/library/long/dm-1104purescalegdpc/

[8] Kyne, F., Clitherow, D., Pimiskern, U. and Schindel, S.: GDPS Family:An Introduction to Concepts and Capabilities, IBM Redbools, http://www.redbooks.ibm.com/abstracts/sg246374.html

日本アイ・ビー・エム株式会社
システム製品事業システム製品
テクニカル・セールス
ICP アドバイザリー IT アーキテクト

佐藤 龍一 **Ryuichi Sato**

[プロフィール]
1990年、日本IBM入社。 メインフレーム製品のテスト、並列シスプレックス基盤の技術支援を担当した後、2000年よりメインフレーム接続ストレージ製品の技術サポートを経て、2004年よりストレージ製品のプリセールスに従事し、金融業のお客様を中心に全業種のお客様を担当。IBMストレージをよりよくお使いいただくための支援を行っている。 ryuhichi@jp.ibm.com