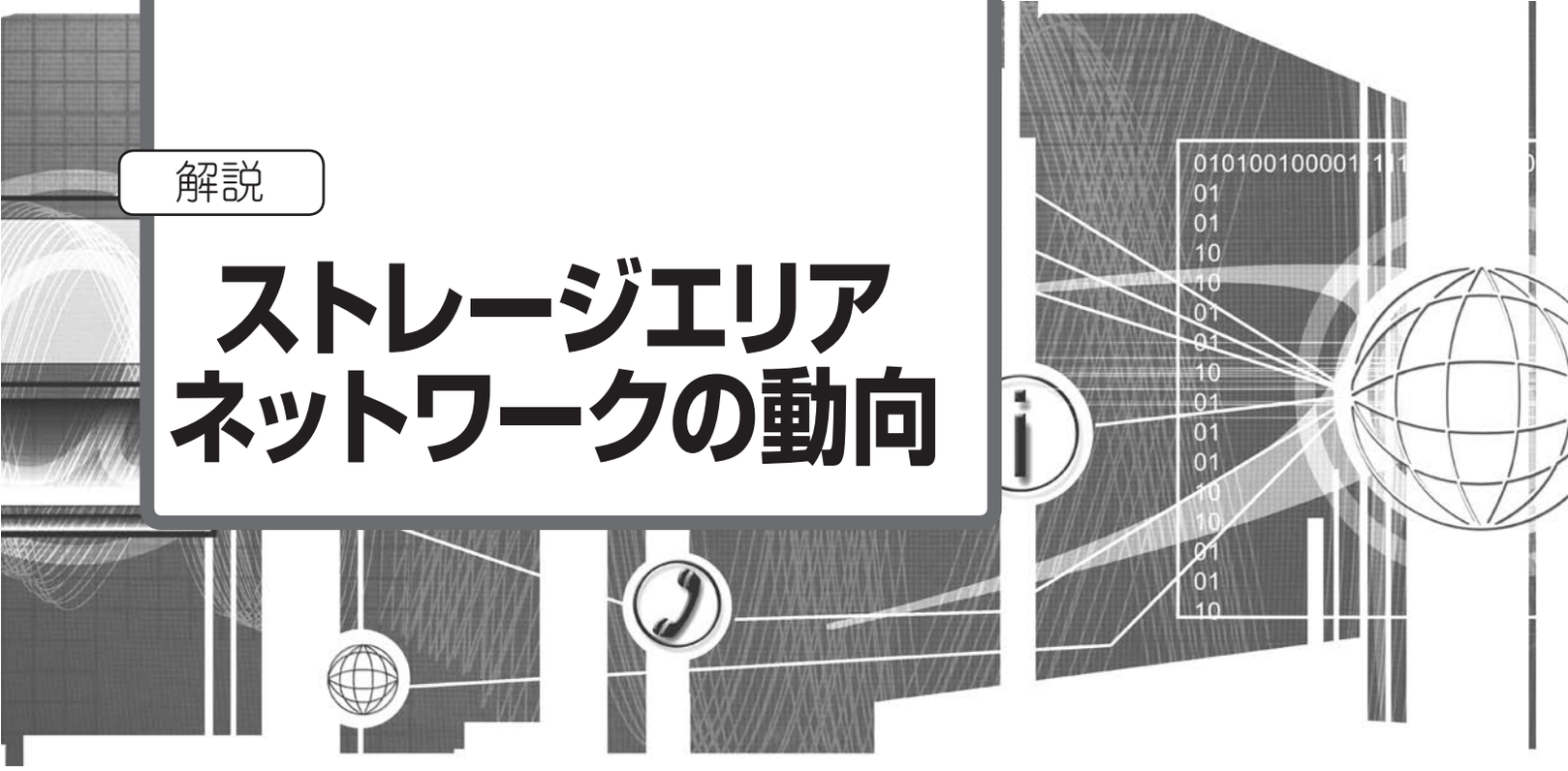

解説

ストレージエリアネットワークの動向

藤田智成　日本電信電話（株）
NTT サイバーソリューション研究所

はじめに

ここ数年，企業や大学が扱うデータ容量は爆発的に増加し，その要求に対応するため，ストレージ技術は進歩している．本稿は，計算機とストレージの接続技術である，ストレージエリアネットワーク技術の最近の動向について解説する．

直接接続型ストレージの限界

計算機とストレージの接続形態は，計算機とストレージを直接接続する Direct Attached Storage（DAS）と，ネットワークを介した接続形態がある（**図 -1**）．ネットワークを介した接続形態は，さらに，Storage Area Network（SAN）と Network Attached Storage（NAS）の 2 種類に分類される．

最も単純な接続形態である DAS では，計算機とストレージが直接接続される．たとえば，パラレル SCSI ケーブルを用いた内蔵ディスクや外付けのディスクアレイとの接続が DAS に分類される．安価なシステムでは，SCSI ディスクの代わりに，ATA やシリアル ATA の内蔵ディスクが用いられることも多い．

DAS は，導入コストは低いが，データ容量や計算機台数の増加に伴い，その欠点が明らかになる．まず，拡張性が計算機に直接接続できるストレージの台数に制限される．また，ディスク容量に空きがある計算機があったとしても，容量不足となった計算機にディスクを追加しなければならないため不経済である．加えて，それぞれの計算機で，バックアップなどのデータ管理を実行しなければならず，ソフトウェアライセンスや人的コストが計算機の台数に比例して増加する．

DAS が今日のストレージに求められる要求に答えることは難しく，ストレージ接続形態の主流は SAN と NAS に移りつつある．

ネットワークを介したストレージ接続

SAN は，DAS で使われる SCSI ケーブルの代わりに，高速なネットワークで計算機とストレージを接続するという考えから生まれた．

現在，SAN で使われる主流のネットワーク技術は Fibre Channel である．計算機は Fibre Channel 用の Host Bus Adapter（Fibre Channel ネットワークのインタフェースカード）を介して，Fibre Channel スイッチに接続され，さらにスイッチとストレージが接続される．通常，各機器は光ファイバで接続される．

SAN と DAS 環境では，計算機とストレージを接続する物理媒体が，光ファイバか SCSI ケーブルかという違いがあるが，計算機からは SCSI ディスクが見えるという点で違いはない．したがって，DAS 環境と同じく，SAN 環境でも，計算機上のアプリケーションは，ext3（Linux）や NTFS（Windows）のようなローカルファイルシステムを介して，または，一部のデータベースのように，ブロックデバイスとして直接，SAN 用ストレージにアクセスする．

商用ベンダから販売されている SAN 用のストレージ

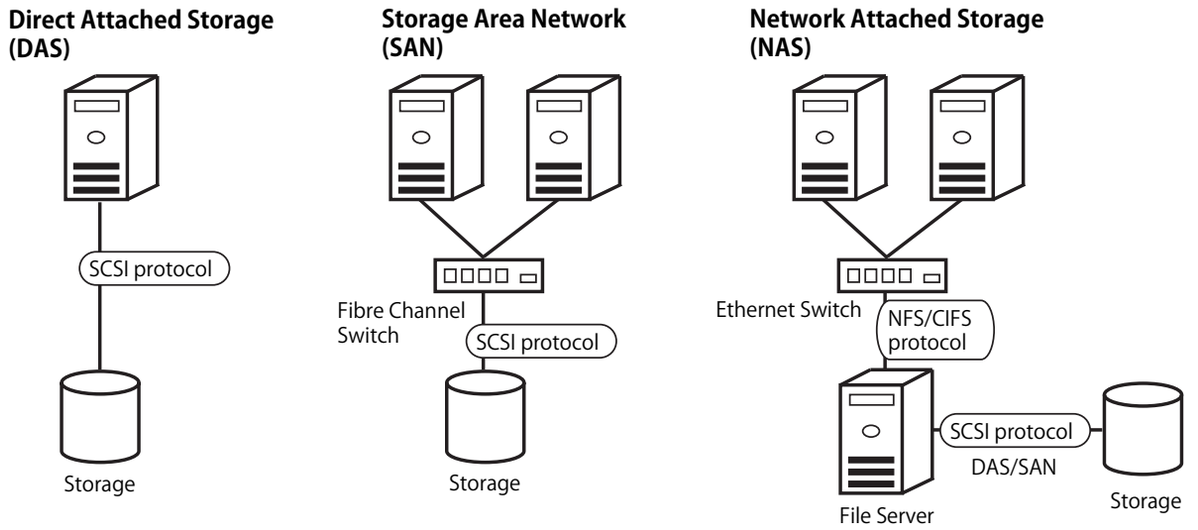

図 -1　DAS・SAN・NAS

は，複数のディスクドライブを搭載したインテリジェントな装置である．ストレージ内部にボリュームマネージャ機能を備え，搭載している物理ディスクドライブを複数の仮想ディスクドライブに分割し，それぞれの計算機に割り当てることができる．SAN は，計算機とストレージの柔軟な構成を可能とし，拡張性，および，ディスク共有による使用効率の向上，冗長構成による信頼性の向上，また，ストレージの一元管理による管理コストの削減につながる．

SAN は DAS を発展させた概念であるのに対して，NAS は，LAN 環境で計算機間のファイル共有を実現するという異なる考えから生まれた技術である．NAS 環境では，計算機は，ネットワーク上のストレージをブロックデバイスではなく，NFS や CIFS などのプロトコルを使って，ファイルシステムとしてアクセスする（**図 -2**）．そのため，NAS 用ストレージは，ファイルサーバと呼ばれる．

NAS は，Fibre Channel よりも低価格なイーサネットハードウェアを使うためコストが低く，計算機間のファイル共有を容易に実現できるという利点を持つ．通常，SAN 環境では，複数の計算機が同じストレージシステムにアクセスするが，それぞれの計算機には個別の仮想ディスクが割り当てられており，データの共有はできない．前述のローカルファイルシステムは，複数の計算機が同時にディスクにアクセスすることを想定していないため，SAN 環境でデータ共有を実現する場合は，計算機はローカルファイルシステムではなく，特殊な SAN 用ファイルシステムを使う必要がある．

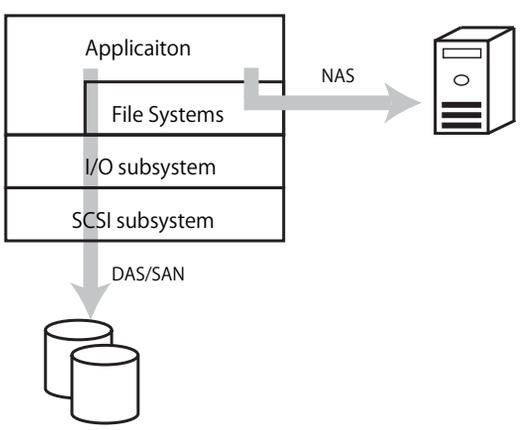

図 -2　ブロックアクセスとファイルアクセス

NAS の欠点としては，一部のデータベースなど，ブロックアクセスを前提として設計されたアプリケーションが動作しない，データ転送がファイルレベルであり，加えて，ブロックレベルよりも上位のファイルレベルのプロトコル処理のため，性能面やスケーラビリティが SAN よりも劣る，などがある．

SAN と NAS は相反するものではなく，適材適所に導入される．また，NAS のファイルサーバのストレージを SAN で接続するなど，両方の利点を組み合わせた構成も用いられる．

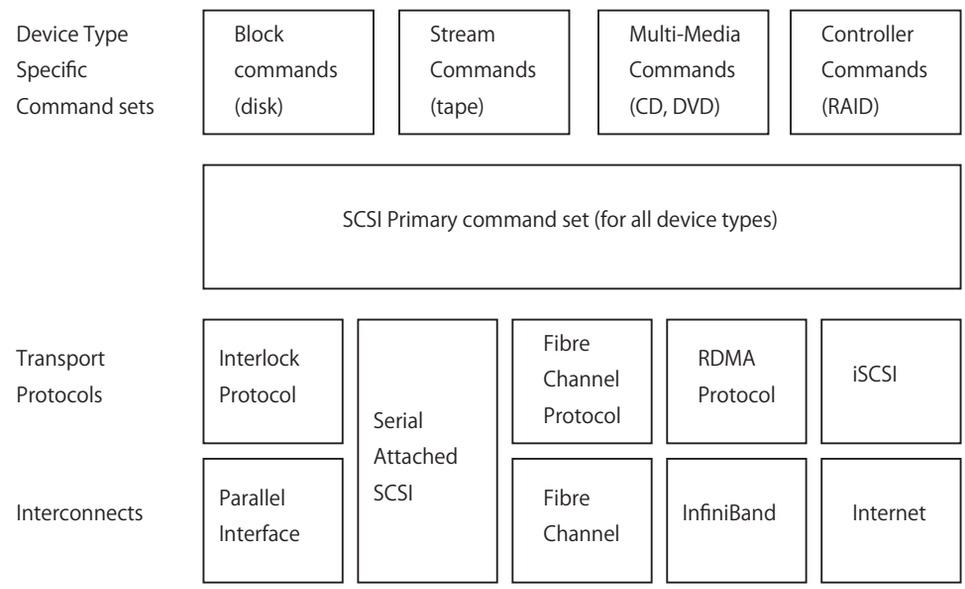

図 -3　SCSI-3 プロトコル

ストレージエリアネットワーク

元々，SCSI 規格は，計算機とストレージをパラレルバスインタフェースを使って接続するために設計されたが，ストレージ技術の進歩とともに，その目的も変化してきた．SAN の登場により，SCSI 規格は，**図 -3** に示すように，デバイス特有のコマンド，基本コマンドセット，トランスポートという，それぞれ独立した階層構造に分割された．SCSI コマンドは，計算機とストレージの接続に用いられるインターコネクト技術とその通信プロトコルから独立しており，Fibre Channel のほかにも，さまざまなネットワーク技術を用いることが可能である．

SCSI RDMA Protocol（SRP）は，RDMA 技術を利用して，SAN を構築するためのプロトコルである．RDMA は，計算機のメモリのデータを，ネットワーク上の他の計算機のメモリに直接転送することができる．RDMA が使えるインターコネクト技術は数種類あるが，計算機とストレージ間の RDMA コネクション確立などのインターコネクト固有な操作と SRP のマッピングが規格化されているのは Infiniband のみである．SRP は，後述する仮想化技術での適用を除けば，Infiniband 同様，ニッチ技術といえる．

現在，最も市場が拡大している SAN 技術は，インターコネクトとして TCP/IP を利用して，計算機とストレージデバイスを接続する iSCSI プロトコルである．iSCSI については，次章で詳しく説明する．

計算機で動作するオペレーティングシステムの SCSI サブシステムの役割は，ファイルシステムやアプリケーションからの I/O 要求を，I/O サブシステム経由で受け取り，SCSI コマンドとしてストレージに送り，その結果を受け取り，適切に処理することである．SCSI 規格の進化に応じて，SCSI サブシステムの設計も大きく変化してきた．たとえば，Linux の SCSI サブシステムは，さまざまな種類のトランスポートを容易に扱えるように，SCSI 規格同様，階層構造に構成されており，インターコネクト固有の機能を提供するソフトウェアコンポーネントはトランスポートクラスと呼ばれている．たとえば，各ベンダの Fibre Channel 用 HostBus Adapter（HBA）ドライバは，Fibre Channel 用トランスポートクラスを利用することで，その実装を簡素化することが可能である．

SAN 環境に対応するため，OS に追加された他の機能としては，ターゲットモードのサポートがあげられる．SCSI プロトコルは，SCSI コマンドを発行する側をイニシエータと呼び，SCSI コマンドを処理する側をターゲットと呼ぶ．通常，計算機がイニシエータ，ストレージがターゲットであり，SCSI サブシステムはイニシエータモードの機能を提供している．ターゲットモードは，商用の SAN 用ストレージの内部で動作している専用のオペレーティングシステムに実装されている機能で，イニシエータから SCSI コマンドを受け取り，必要に応じてディスクドライブのデータ入出力を実行し，適切な結果を送る機能を有する．計算機とストレージを接続するインタフェースである HBA は，一般に，イニシエータとターゲット両方のモードをサポートしており，Linux のような汎用 OS でも，SCSI サブシステムがターゲット

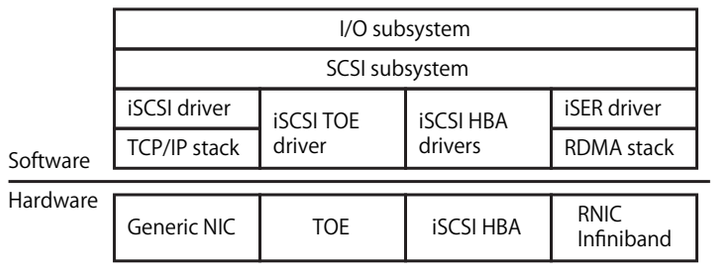

図 -4　iSCSI で利用可能なインターコネクト技術

モードをサポートすることで，商用ストレージと同様の機能を実現することが可能である [1)]．

本稿では，計算機とストレージをネットワークを介して接続し，計算機がストレージにブロックレベルでアクセスする技術を SAN と呼ぶ．SCSI 以外のブロックレベルのプロトコル，たとえば，ATA プロトコルをイーサネットのフレームにのせる ATA Over Ethernet（AOE）も広義の SAN 技術といえる．しかし，ATA プロトコル自体が，ネットワークを介した接続を想定していないため，システムバスよりも，通信条件が悪いインターコネクトに必要なエラー処理が考慮されておらず，その適用範囲はブレードサーバ内部などに限られている．

SAN で利用可能なさまざまなネットワーク技術のうち，どれを選択するかは，コスト，接続距離，帯域などの要素に依存する．接続距離と帯域の条件は，計算機で動作させるアプリケーションの要求に依存し，絶対的な指標はない．

接続距離が長くなるほど，ストレージシステムから SCSI コマンドの応答が返ってくるまでの遅延時間が長くなる．通常のディスクアクセス時間が数 ms であることを考慮して，アプリケーションに影響を与えない遅延になるようにしなければならない．

帯域に関しては，Fibre Channel で一般的に使われている速度は，1，2，または，4 Gbps である．iSCSI では，ギガビットイーサネットが使われることが多いが，ハイエンドなシステムでは，10Gb イーサネットも使われる．

iSCSI プロトコル

DAS から，Fibre Channel を使った SAN への移行を阻む最も大きな要因は，そのコストにあると考えられてきた．Fibre Channel HBA，Fibre Channel スイッチは，他のインターコネクト技術のハードウェアと比較して高価である．また，Fibre Channel の導入，運用には高度な専門知識が要求され，そのような知識を持った人員が少ない点も問題となっている．

iSCSI は，インターコネクトとして TCP/IP を利用する SAN 技術である．イーサネット機器を利用することで，安価に SAN を導入することができる．さらに，既存のネットワーク管理アプリケーションを使って，SAN トランスポートを管理できる．また，イーサネットと TCP/IP に精通した管理者は多いため，iSCSI 環境運用に必要な知識は，Fibre Channel よりも容易に習得することができる．

Fibre Channel は，専用のネットワークを使うこともあり，セキュリティをあまり考慮していない．ストレージシステムが提供する HBA の識別子を使ったベンダ独自のアクセスコントロール機能が使われることが多い．一方，iSCSI の場合，他のアプリケーションとネットワークを共有する構成や公衆網を使った広域環境での運用が多いことを考慮して，セキュリティに配慮した設計がされている．計算機とストレージ間の CHAP 認証などのさまざまな認証機能，IPsec を使った計算機とストレージ間のセキュアな通信などの機能が標準化されている．加えて，一般に，商用のストレージシステムは，IP アドレスをベースにしたアクセスコントロールなど，独自のセキュリティ機能を提供する．

iSCSI で用いることのできるインターコネクトは 4 種類に分類される（**図 -4**）．

第 1 の接続方法は，通常の NIC を用いて，OS が iSCSI プロトコルを処理する方法である．HBA ドライバとして実装された iSCSI イニシエータドライバが，TCP/IP スタックを利用し，SCSI コマンドをストレージに送る．iSCSI イニシエータドライバは，OS の TCP/IP スタックを経由して，NIC にアクセスするため，すべてのベンダの NIC が利用できる．

現在，大半の計算機はギガビットの NIC を標準搭載しているため，第 1 の接続方法はコストが低いという利点がある．しかし，iSCSI プロトコル処理，および，TCP/IP プロトコル処理を計算機の CPU で実行するため，計算機の負荷が上昇する．加えて，冗長なメモリコピーによって，メモリバス帯域を消費する点も問題である．READ コマンドを処理する際，iSCSI ドライバは，TCP のペイロードとして送られてくる iSCSI ヘッダとストレージのデータを，一時的なバッファに保存してから，iSCSI ヘッダを処理して，データの最終的なバッファ（ファイルシステムのページキャッシュやアプリケーションが指定したバッファ）を判断して，データをコピーする必要がある．Fibre Channel プロトコルや SRP では，HBA は，メモリコピーを伴わず，受け取ったデータを最終的なバッファに直接届けることができる．これらの問題は，10Gb イーサネットで特に問題となる．

第 2 の接続方法は，TCP Offload Engine（TOE）と呼ばれる NIC を使う方法である．TOE は，通常の NIC に TCP/IP プロトコル処理のためのハードウェアを追加したもので，計算機の CPU 負荷を軽減することができる．ただし，TOE は iSCSI プロトコル処理はできないので，上記のメモリバス帯域を消費する問題は解決できない．

第 3 の接続方法は，商用ベンダの iSCSI HBA を使う方法である．iSCSI HBA は，TOE に iSCSI プロトコル処理のためのハードウェアを追加したものであり，計算機の CPU から，iSCSI と TCP/IP プロトコル処理をオフロードすることができる．かつ，iSCSI HBA が iSCSI ヘッダを処理するため，データを直接，最終的なバッファに保存することができる．この接続方法の欠点は，ハードウェアコストが高い，および，各ベンダの iSCSI HBA 専用のドライバがそれぞれ必要になる点である．

第 4 の接続方法は，Infiniband などの RDMA インターコネクトと，データ転送に RDMA を利用するように iSCSI を拡張した，iSCSI Extensions for RDMA Specification（iSER）プロトコルを利用する方法である．

iSER が期待されている理由の 1 つは，イーサネットに RDMA 機能を追加する Internet Wide Area RDMA Protocol（iWARP）の存在である．これまで，RDMA が使えるさまざまなインターコネクト技術が開発されてきたが，いずれの技術も，イーサネットのような大きな市場を得るには至っていない．iWARP は，既存のイーサネットワークインフラストラクチャや管理者の専門知識を有効に活用できる技術として期待されている．

iWARP をサポートするハードウェアインタフェースは，RDMA NIC（RNIC）と呼ばれ，TOE に RDMA 機能のためのハードウェアを追加した NIC である．すでに，複数のベンダの RNIC が利用可能である．

OS の設計によって異なるが，Linux のように，Infiniband や RNIC などの RDMA インターコネクトにアクセスするための共通 API を実現している OS では，1 つの iSER 用のドライバが，その API を通じて，すべてのベンダの RDMA インターコネクトを扱うことができる．

TCP/IP は複雑なプロトコルであり，TOE のように，TCP/IP をハードウェアで実装することは容易ではない．簡素化のため，IPv6 や IPsec などの機能を実装していないハードウェアインタフェースもある．

ディザスタリカバリ

SAN は，DAS を置き換えるだけでなく，その利点を活かしたソリューションとしても用いられる．代表的なソリューションの 1 つが，遠隔地へのデータのバックアップである．自然災害やテロ事件などが，相次いで発生しているが，現在の企業は，業務拠点が失われるような大規模な災害が発生した場合でも，データの損失を防ぎ，迅速に業務を再開しなければならない．この要件を実現するのがディザスタリカバリと呼ばれるソリューションであり，遠隔地へのデータのバックアップはリモートミラーリングと呼ばれる．ディザスタリカバリソリューションでは，**図 -5** に示すように，通常時に業務を行うローカルサイトのストレージシステムを，リモートサイトのストレージシステムと接続し，データを同期する構成がとられる．ストレージシステム同士でデータを同期するため，計算機への影響が小さい．

Fibre Channel を使ったリモートミラーリングは，ストレージシステム間を直接光ファイバで接続するか，Internet Fibre Channel Protocol や Fibre Channel over IP と呼ばれる，Fibre Channel プロトコルを IP パケットにカプセル化するプロトコルを実装したゲートウェイ装置を両拠点に設置して，IP ネットワークを利用する．Fibre Channel と比較すると，iSCSI は，公衆回線や広域 LAN などのサービスを利用して，低コストでリモートミラーリングが実現できるため，その導入に拍車をかけている．

同期型のリモートミラーリングでは，ローカルとリモートサイト，両方のストレージシステムのデータを更新してから，計算機に SCSI コマンドの完了を通知する．この方法は，両方のストレージシステムのデータが常に同期しているという利点があるが，ストレージシステム間のネットワーク遅延が，ローカルサイトの計算機で動作するアプリケーション性能に影響するという欠点がある．そのため，ストレージシステム間の距離が制限され，一般には，専用線などの高品質のネットワークが必要とされる．

ネットワークの遅延が大きい環境では，非同期型のリ

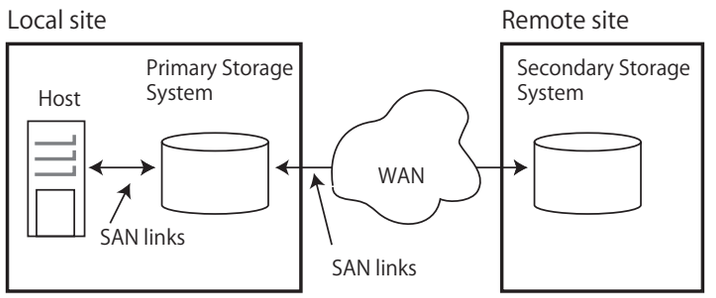

図 -5　ディザスタリカバリの構成例

モートミラーリングが使われる．非同期型は，ローカルサイトのストレージシステムのデータを更新すると，計算機に SCSI コマンドの完了を通知し，後で，ローカルサイトのストレージシステムから，リモートサイトのストレージシステムにデータを転送する．災害でローカルサイトのストレージシステムが失われた場合に，リモートサイトのストレージシステムに同期されていない最新のデータが失われてしまう可能性があるが，アプリケーションの性能がネットワーク遅延の影響を受けない．

仮想化技術

1 台の計算機を複数の計算機に論理的に分割し，複数の OS を同時に動作させる仮想化技術も，ディザスタリカバリ同様に，SAN をキー技術として使う技術の 1 つである．ストレージと同様に，IT システムの肥大化，複雑化に対応するために計算機の台数が増加し，その運用，管理に必要な設備，人的リソースに伴うコストの増加が問題になっているため，仮想化技術は注目を浴びている．仮想化技術を使うことで，1 台の計算機で，Linux，FreeBSD，Windows，Solaris などの異なる OS を動作させることができるので，複数の計算機のリソースを 1 台の計算機に統合することが可能となる．

計算機に仮想化環境を実現するソフトウェアは，仮想マシンモニタ（Virtual Machine Monitor）と呼ばれ，計算機全体を制御し，CPU，メモリ，ストレージなどの論理的に分割された計算機のリソースから構成される論理的な計算機，仮想マシンで OS を動作させる．仮想マシンで動作する OS はゲスト OS と呼ばれる．

仮想化技術と SAN を組み合わせ，複数の計算機でストレージを共有しているシステムでは，計算機間での仮想マシンの移動[☆1]などの高度な運用が可能である．このようなシステムでは，計算機に搭載された SAN インタフェース（Fibre Channel HBA，NIC など）を介して，複数の仮想マシンが SAN ストレージにアクセスする（**図 -6**）．

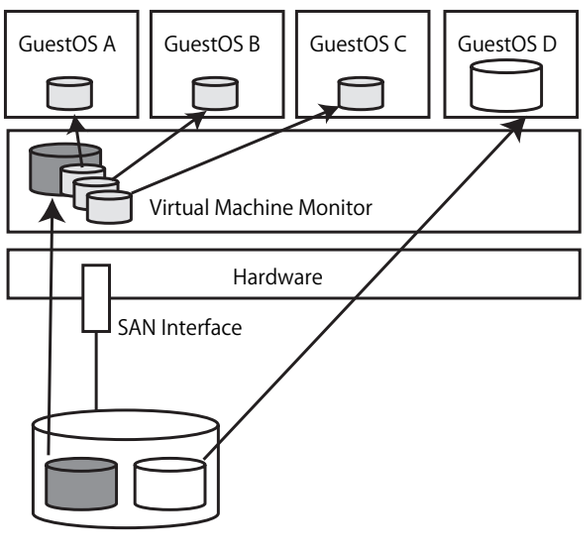

図 -6　SAN と仮想化技術の組合せ

仮想化環境でのストレージの構成は，仮想マシンモニタが論理的に分割する方法と，SAN ストレージの仮想ディスクをそのまま使う方法の，大きく 2 種類に分けられる．

図 -6 中のゲスト OS の A，B，および C には，仮想マシンモニタが仮想ディスクを論理的に分割する方法が適用されている．SAN ストレージから見ると，仮想マシ

☆1　ダウンタイムなしのハードウェアアップグレード，計算機間での負荷分散などの目的で使われる．

ンモニタのみがストレージにアクセスするため，1 台の計算機に仮想ディスクを提供するという従来の環境との変化はない．仮想マシンモニタは，その仮想ディスクを論理的に分割し，それぞれのゲスト OS に提供する．仮想マシンモニタとゲスト OS の間のストレージインタフェースは，仮想マシンモニタの実装によって異なる．

IBM の Power アーキテクチャでは，仮想マシンモニタとゲスト OS 間のストレージインタフェースとして SRP プロトコルを用いる [2]．ゲスト OS からは，仮想マシンに SRP 用 HBA が搭載されているように見える．仮想マシンモニタのかわりに，VIO サーバと呼ばれる 1 個の特別な仮想マシンが仮想ディスクの論理的な分割を担当する．VIO サーバでは，ゲスト OS として，Linux または専用 OS が動作し，残りの仮想マシンと RDMA 通信を行い，SRP ストレージと同等の機能を提供する．通常，VIO サーバは SAN ストレージの仮想ディスクをローカルファイルシステムとして使用し，作成したファイルを，残りの仮想マシンに SCSI ディスクとして提供する．

オープンソースの仮想マシンモニタである Xen[3] では，Power アーキテクチャ同様に，特別な仮想マシンが，仮想ディスクの論理的な分割を担当するが，仮想マシン間のストレージインタフェースは独自のブロックプロトコルを利用する．SCSI プロトコルと比較すると，Xen 独自のプロトコルは単純であるが，ゲスト OS の既存のフレームワークにうまく適合しない，デバイス管理機能が十分ではない，などの問題がある．Xen でも，Power アーキテクチャ同様，SRP プロトコルを使ったストレージインタフェースが開発されている [4]．

図 -6 のゲスト OS の D には，SAN ストレージの仮想ディスクをそのまま使う方法が適用されている．仮想マシンモニタによって，仮想マシンには，HBA が搭載されているように見える．仮想マシンモニタは，ゲスト OS が発行した SCSI コマンドを，実際の HBA を通じて，SAN ストレージに直接転送する．仮想マシンモニタで仮想ディスクを論理分割する方法と比較すると，仮想マシンモニタと連携した高度な制御の実現が難しくなるが，論理分割に伴うオーバーヘッドを削減することができる．また，SAN ストレージの既存の仮想ディスクと保存されているデータを，そのまま仮想マシンに割り当てることができるため，仮想化技術を容易に新規導入できるという利点がある．

それぞれの仮想マシンが直接 SAN ストレージにアクセスする場合でも，仮想マシンモニタは，1 個の SAN インタフェースを通じて，それぞれの仮想マシンが干渉せずに，個別の計算機としてアクセスできるようにする必要がある．仮想マシンモニタを経由することで発生するパフォーマンスの低下を避けるために，SAN インタフェースが，仮想マシン間の干渉を防ぐための機能，たとえば，論理的に複数のインタフェースのように動作するなどの機能を提供することで，仮想マシンが SAN インタフェースに直接アクセスする試みもなされている [5]．

まとめ

本稿では，今日の企業のストレージの主流になりつつあるストレージエリアネットワーク技術の動向について説明した．

従来の直接接続型ストレージが増加し続けるデータ容量に対応できなくなっていることに加え，イーサネットを用いて低コストでストレージエリアネットワークを実現する iSCSI 技術の成熟や，ストレージエリアネットワークがキー技術として使われるディザスタリカバリや仮想化技術を使ったシステムの需要増加から，今後，ストレージエリアネットワークの導入がさらに加速することが期待される．

参考文献

1）Fujita, T. and Christie, M.: tgt: Framework for Storage Target Drivers, Ottawa Linux Symposium, pp.303-312(2006).
2）Boutcher, D. and Engebretsen, D.: Linux Virtualizationon IBM POWER5 Systems, Ottawa Linux Symposium, pp.113-120 (2004).
3）Barham, P., Dragovic, B., Fraser, K., Hand, S., Harris, T., Ho, A., Neugebauer, R., Pratt, I. and Warfield, A.: Xen and the Art of Virtualization, The 19th ACM Symposium on Operating Systems Principles, pp.164-177 (2003).
4）Fujita, T.: Xen Scsifront/back Drivers, Xen Summit (2006).
5）Liu, J., Huang, W., Abali, B. and Panda, D. K.: High Performance VMM-Bypass I/O in Virtual Machines, the USENIX Annual Technical Conference, Boston, MA, pp.29-42 (2006).

（平成 18 年 12 月 4 日受付）

藤田智成（正会員）
fujita.tomonori@lab.ntt.co.jp

2000 年早稲田大学大学院理工学研究科修士課程修了．同年日本電信電話（株）入社．オペレーティングシステム，ストレージシステムに関する研究に従事．ACM, USENIX 各会員．