IP ネットワーク接続ストレージソリューション

IP-Network Attached Storage Solutions

根 来 元

要 約 ブロードバンドネットワークでは，大量のデータとリッチコンテンツに対する高速でかつ高品質なアクセスの要求が日々増大している．フラットファイル，Web コンテンツ，データベースといったあらゆるタイプのデータが，作成者からエンドユーザへと配信される．このためブロードバンドネットワークシステムでは，Web 層，アプリケーション層，データベース層の各層にスケーラブルで可用性にとんだ IP ネットワーク接続ストレージが必要になる．IP ネットワーク接続ストレージアーキテクチャでは，ストレージアプライアンスと IP ネットワークインフラストラクチャが高度に統合され，高速ユニバーサルアクセスとマルチベンディングな相互接続を実現する．本稿では，第 2 章で一般的な IP ネットワーク接続ストレージソリューションを紹介した後，大容量でバラエティにとんだデータを配信するブロードバンドネットワークに対応するために考えられたさまざまな IP ネットワーク接続ストレージ関連の新技術を紹介する．

Abstract In a broadband network, the requirement for fast, and high-quality access to large volume data and rich contents is expanding. All type of data, ranging from flat files, Web contents and database, needs to be distributed from creators to end-users. Broadband network applications require the highly scalable and available IP-network attached storage that delivers exceptional performance for each tier of Web, application, and database. The IP-network attached storage architecture sophisticatedly integrates storage appliances and IP-network infrastructure, and enables high-performance universal accesses and interconnections. The second chapter of this paper describes four typical solutions about IP-network attached storage. The subsequent chapters introduce several new technologies of IP-network attached storage to implement the broadband network that can distribute large volume and various kinds of data to clients.

1. はじめに

本稿では，ブロードバンドネットワークの不可欠な構成要素である IP ネットワーク接続ストレージ技術に焦点をあてる．IP ネットワーク接続ストレージソリューションでは，ストレージアプライアンスと IP ネットワークインフラストラクチャが高度に統合され，高速ユニバーサルアクセスとマルチベンディングな相互接続を実現する．その結果として，エンタープライズレベルでのストレージアクセス，クラスタリング，リモートミラー，バックアップなどの機能が IP ネットワークを通して提供され，データセンタ内はもちろん CAN，MAN，WAN にまたがるスケーラブルなストレージインフラストラクチャの構築が可能になる．IP ネットワーク接続ストレージをクラスタ構成にすれば，高いデータ可用性が保証できる．さらにストレージ管理ソフトウェアを導入すれば，グローバルな管理機能やリモートなバックアップ/リストア，および災害対策などのソリューションを実現することもできる．

IP ネットワーク接続ストレージとは，NFS（Network File System）や CIFS（Common In-

ternet File System) といったオープンな標準ファイルレベルインタフェースを行う NAS (Network Attached Storage) やオブジェクトレベルのアクセスを行う http や ftp をさす．IP ネットワークを通じてブロックレベルインタフェースを提供する iSCSI (Internet SCSI) プロトコルをはじめとする IP-SAN も，ごく近い将来にこの一部として含まれることになる．

2. IP ネットワーク接続ストレージソリューション

なぜ IP ネットワークにストレージを接続するのかといえば，グローバルなネットワークがすでに IP および Ethernet に収束しており，その上にストレージインフラストラクチャを構築することが容易だからである．Gbit Ethernet そして 10 Gbit Ethernet が登場し，SCSI や Fibre Channel を上回るペースで高速化も進んでいる．ストレージインフラストラクチャ構築のために，膨大なコストをかけて IP ネットワークとは別のネットワークインフラストラクチャを新たに構築する理由は何もないのである．ここでは，IP ネットワーク接続ストレージによる現在の代表的なソリューション例を示す．

2.1 運用管理コスト削減――e-ビジネスアプリケーション

個人的なインターネットポータル，B 2 B アプリケーション，企業イントラネットアプリケーションといった e-ビジネスアプリケーションでは，Web 層，アプリケーション層，データベース層によって構成される階層型アーキテクチャの各層にスケーラブルで可用性にとんだ高速 IP ネットワーク接続ストレージが必要である．ところで，Web 層にストレージを配備する方法としては二つ考えられる．

一つは，各 Web サーバに接続された DAS (Direct Attached Storage) に Web コンテンツを格納する方法である．この方法ではコンテンツの冗長コピーを Web サーバごとに持つことになり，ストレージリソースを浪費する．さらに各 Web サーバ上のコンテンツを更新/管理するのが難しく，全 Web サーバ上のコンテンツを変更するのは大変である．ストレージ容量を増大するために頻繁にサーバをアップグレードしなければならず，Web サーバファームの拡張性という点で非効率的である．もう一つの方法は，全コンテンツを IP ネットワーク接続ストレージに統合する方法である．全 Web サーバがこの共有リソースを利用するため，個々のサーバにデータの複製を置く必要はない．全データが統合ストレージに格納されるため，コンテンツ管理の負荷も軽く，障害発生時や保守時にも Web サーバを簡単に交換することができる．e-ビジネスアプリケーションは 24 時間/365 日利用できなければならない．図 1 は，可用性が高く耐障害性に優れた e-ビジネスアプリケーション環境を示している．

2.2 高可用 (HA) システム構築――データセンタシステム

企業のデータセンタで展開される ERP (Enterprise Resource Planning)，CRM (Customer Relationship Management)，SCM (Supply Chain Management) といったビジネスアプリケーションは，非常に可用性の高いネットワーク型ストレージインフラストラクチャを必要とする．企業がインターネット戦略を導入すると，最終的には必ずこれらの Web 対応ソリューションに発展する．

高可用システムは，サーバ，ネットワークおよびストレージのデザインに冗長性を持たせることで実現する．図 2 は，クラスタ構成の IP ネットワーク接続ストレージと各種サーバを Gbit

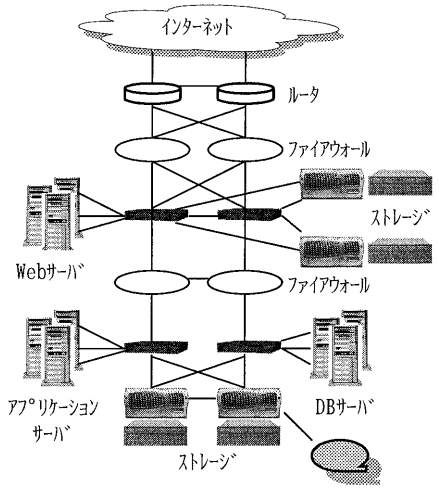

図 1　e-ビジネスアプリケーション

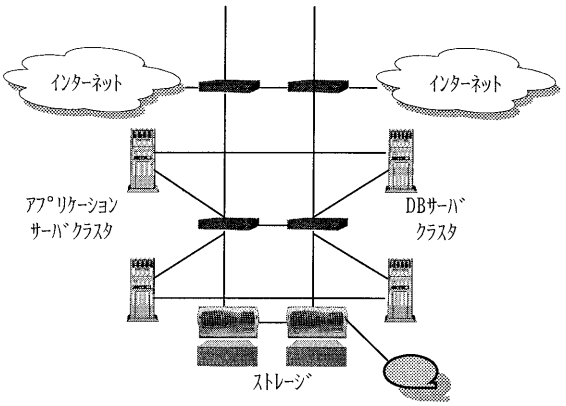

図 2　高可用システム

Ethernet に配備して ERP アプリケーションをサポートするデータセンターの例である．ネットワーク帯域幅を効率的に使用するために，サーバは Ethernet サブネットのトラフィックを制限するようにスイッチ上で構成されている別個の VLAN に接続される．NAS はバックエンド上の各自の VLAN に置かれ，スイッチはこれらにあてられたトラフィックをそのサブネットにルーティングする．

2.3　高速アクセス――ワークグループコラボレーション

ソフトウェア開発やエンジニアリングデザイングループなどのエンジニアリング環境では，管理上のオーバヘッドを抑え，かつ高速 LAN 上でのデータ共有が可能なインフラストラクチャが必要である．こうしたアプリケーションとして，EDA シミュレータや合成ツール，CAD/CAM/CAE のツール，ソースコード管理，バージョン管理などがあげられる．

図 3 では，冗長構成のスイッチを通して各種のクライアントがクライアントアクセス層で集約される．分散層のスイッチに接続する冗長 Gbit Ethernet アップリンクは，代替パス，リンク障害時の高速コンバージェンス，および冗長スイッチ上の負荷分散を提供し，単一モジュール構成の障害による機能停止を防ぐ．ストレージも，障害に備えてクラスタ構成される．ストレージは，分散層でマルチレイヤスイッチの Gbit Ethernet ポートに接続される．分散層はまた，ワークグループが使用するアプリケーションがこの層でアプリケーションサーバを要求した場合に，アプリケーションサーバへの接続を提供する．これによって，エンジニアリングアプリケーションを実行するアプリケーションサーバを IP ネットワーク接続ストレージに接続するためのプライベートな専用ストレージネットワークが形成される．可用性をさらに高めるためには，分散レイヤスイッチも冗長構成に配備する．

2.4　セキュアな WAN――分散型エンタープライズストレージ

各地に分散して拠点を持つ企業では，効果的なコラボレーションと災害復旧のための分散型ストレージネットワークが必要となる．通常こうしたリモート拠点は，ハブおよびスポークトポロジによって，相互にまた中央拠点に接続しているため，リンク上でデータを交換するため

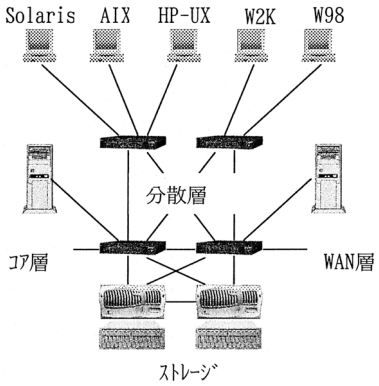

図 3　高速アクセス

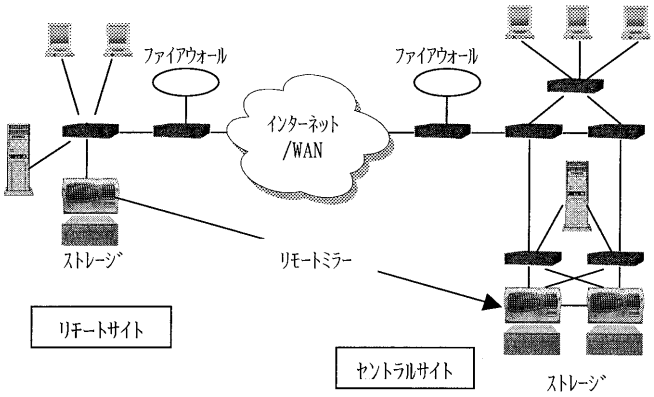

図 4　セキュアな WAN

のセキュアな WAN 環境を必要とする．ストレージネットワークは，WAN 帯域幅を安全かつ効率的に使用できるようにデザインされる必要がある．また，リンク費用から高い ROI（対投資利益率）を生み出すために，WAN リンクを効率的に使用する必要がある．この WAN 接続は，中央拠点インフラの WAN モジュールに接続するフレームリレーや ATM などの WAN サービスを使った専用リンクを通して提供される．代わりに既存のデータリンク上でセキュアな VPN（仮想閉域網）接続を展開することもできる．図 4 は，こうした要件に対処するための分散型エンタープライズストレージ環境の展開例を示している．

図 4 に示されているスイッチやルータは，各種 WAN モジュール，ファイアウォールセキュリティ，およびハードウェアによる暗号化を行い，VPN 接続を実現する．また WAN リンクのトラフィック輻輳の問題に対処する各種 QoS 機能も装備されている．これらの機能によって，WAN リンク帯域幅がストレージ関連のトラフィックによって適切に使用されるようになり，他のミッションクリティカルなアプリケーションデータ用のリソースが枯渇することがなくなる．各地に分散したグループ間でコラボレーションを行う場合は，各拠点でデータを共有する．

3.　CDN（Content Delivery Network：コンテンツ配信ネットワーク）

3.1　CDN とは何か

CDN とは，クライアントに対してコンテンツをより効果的に配信できるように構成されたネットワークおよびストレージの集合体である．通常，第 1 層としての Web 層，第 2 層としてのアプリケーション層，第 3 層としてのデータベース層の 3 階層で構成される．特に第 1 層は，LAN/WAN との接点，つまりエンタープライズシステムとネットワークの先端部分であることから“エッジコンピューティング”とも呼ばれる．エッジでコンテンツをキャッシュすることで，配信パフォーマンスが向上して帯域幅の使用量が減少し，CDN の TCO（総所有コスト）を削減することができる．ここでは，CDN における”Center-to-Edge アーキテクチャ”で Center に位置するバックエンドストレージと Edge に置かれるエッジサーバについて述べる．

3.2 バックエンドストレージ

CDN バックエンドストレージの第一の役割は，コンテンツをステージングすることにある．コンテンツの開発環境で CIFS/NFS のマルチプロトコルがサポートされていれば，Windows PC と UNIX ワークステーションにまたがったコンテンツのステージングが可能になり，同一データを両方から共有することができる．また，一次サイトの Web サイトのストレージにステージングされたコンテンツをリモートミラーリングによって，別の Web サイトのストレージに複製し，それを複製先の二次 Web サイトのバックエンドストレージとすることもできる．ユーザはもっとも近い複製にリダイレクトされる．一次サイトへの接続が失われれば，代わりの複製にユーザをリダイレクトした後に災害復旧を行えばよい．中央のステージングサーバからローカル/リモートの Web サーバやエッジまでのコンテンツの配信は，ポリシーベースで完全に自動的に行われ，事前に配信することもできる．このポリシーベースの管理は，ブラウザベースのインタフェースを通して簡単に行える．ジョブステータスの自動報告，コンテンツのバージョン管理，ロールバック復元機能などが提供される．

したがって，CDN コンテンツ管理用バックエンドストレージは，垂直にも水平にも優れたスケーラビリティを持っていなければならない．システム作動中にシェルフ/ディスクを追加でき，多数の大容量ストレージを既存のネットワークに速やかに追加できなければならない．また，e-ラーニングのようなサービス品質が重視されるアプリケーションのために，ストリーミング CDN が必要とする高パフォーマンスを低い TCO で提供できなければならない．さらに，クラスタフェイルオーバーによって互いのデータを引き継ぐように複数ストレージを連携させるべきである．クラスタ化されたシステムをミッションクリティカルなアプリケーションのバックエンドに使用すれば，単一個所からの障害の発生を防止することができる．

3.3 エッジサーバ

ユーザからのアクセスを直接受けるのが第 1 層の Web 層，つまりエッジである．システムの外側から見れば，エッジはサイトへの入り口であり，内部構造を隠蔽するための外殻である．一方，システムの内側から見ると，ユーザからのアクセスをアプリケーションサーバやデータベースに転送するフィルタのような存在である．つまり，LAN と WAN の接点をエッジと呼ぶ．具体的には，Web サーバ，FTP サーバ，メールサーバなどを指すが，広義にはファイアウォールや VPN（Virtual Private Network）装置なども含まれる．エッジサーバの目的は，これまでバックエンドに置かれていた業務ロジックやデータをエンタープライズシステムのエッジにキャッシュし，負荷分散やルーティングをより効率的に行うことである．Web ベースのコンテンツのほとんどはデータベース駆動型で，動的に生成されるが，伝送されるコンテンツの大部分は，頻繁には変更されない静的なデータである．こうしたデータをエッジに移すわけである．アプリケーションやデータをエッジに展開しようという試みの多くは，モバイルデバイスからのアクセスを想定したものである．

典型的な“エッジ分散モデル”は，小規模 CDN で使用される．この場合，エッジをネットワークエッジ，リモートオフィス，POP（アクセスポイント）に設定して発信元の帯域需要を削減する．“エッジ階層モデル”は，多数のハブや POP が存在するより複雑な CDN で使用される．“ハブ・アンド・スポーク・モデル”は，大規模企業ネットワークによって使用される．この場合，発信元のデータは災害復旧用にリモートミラーリングでリモートに複製し，リ

モートハブの小型ストレージのコンテンツを 1 台の大容量ストレージに統合したりする．さらに，そこからテープにバックアップすることもできる．また，Snapshot によってボリューム全体の災害復旧を迅速に行なえるようにする．

エッジコンピューティングは，デバイスを並列構成で配置することで，全体としての処理能力を高める“ 水平方向スケーリング ”の特性を持つ．大量に投入する際の支障とならないように，エッジで使用されるサーバやストレージは，第 2 層や第 3 層に比べると，それほど高い処理能力や RAS（Reliability，Availability，Serviceability）は要求されない．そのため，価格が安価で設置スペースを極力抑えるためにコンパクトな筐体の製品が最適ということになる．とはいえ，クラスタリング機能，冗長構成，NAS への高速アクセス機能は必須である．また，ハードウェアルーティング機能や SSL 機能も組み込まれる必要がある．そこで使用されるのが，キャッシュサーバであり，ブレードサーバである．ブレードサーバは，通常のサーバマシンの機能が実装された回路基盤“ Blade ”を，専用のラックマウント型筐体に複数枚装着することで多数のサーバマシンの高密度実装を実現するアーキテクチャである．さらに最近では，ブレードサーバ用ストレージを組み合わせて“ ブレードラットフォーム ”ともいわれるまでになってきている．ここで使用されるストレージは NAS であるが，小型ながらも SAN と同等なネットワークストレージ環境を実現する．

4.　ニアオンライン大容量ストレージ

4.1　データ保護

ストレージ上にあるデータの保護を考えるとき，2 種類のデータ保護を分けて考える必要がある．一つは，ハードウェア障害，ソフトウェア障害，人為的障害によってデータを紛失してしまった際に，これを“ 復旧 ”することを目的とした“ データバックアップ ”である．もう一つは，将来の参照に備えてアーカイブデータを長期間“ 保管 ”しておく“ データ保存 ”である．この場合，コンテンツによっては数ヶ月，数年，数十年にわたって保存されることになる．これまでは，“ バックアップ ”のためのデータも，“ 保存 ”のためのデータも，等しく磁気テープに保存されてきた．このため，ストレージ上のデータを磁気テープにコピーするためのさまざまな技術が検討され発展してきた．

DLT，LTO，AIT と次々と新しいテープの規格が作成され，シングルドライブからオートローダ付きのライブラリ装置まで対応している．バックアップソフトウェアも Veritas 社，Legato 社，Computer Associates 社などが，争うように自社製品の機能強化を続けている．今やアプリケーションを止めないで行なうオンラインバックアップが，当然のようにユーザから要求される．また，差分バックアップ，増分バックアップなどいろいろなバックアップ方法が，テープのローテーション方法などと組み合わされて考え出されてきた．テープ装置の接続形態も，以前の直付けの形態から，ネットワークバックアップ，SAN（Storage Area Network）による LAN フリーバックアップ，そしてサーバレスバックアップなど，次々と新しいバックアップ形態が登場しつつある．しかし，こうした技術はすべて“ いかに効率よくストレージ上のデータをテープに吸い上げるか ”を追求するためのものである．

近年の急激なデータの増加とデータの重要性の増大に伴い，障害発生時には膨大なデータを非常に短時間で“ 復旧 ”させることが要求されるようになると，“ バックアップ ”を目的としたデータを磁気テープにとることは意味をなさなくなりつつある．現在の最速のテープ装置を

駆使しても，数 TB のデータを復旧するには数十時間から数日を要する．これは 24 時間/365 日稼働が要求されるアプリケーションシステムにおいては，まったく受け入れられない数値である．ごく最近まで，データのバックアップを，“ 復旧 ” するという観点から考えられたことはほとんどなかった．同じことは，突然の参照を前提としたアーカイブデータについてもいえる．気象情報，I-モード映像，医療画像などのデータは，磁気テープはもちろん CD-ROM や DVD に保存してあったにしても，保存データを瞬間的に取り出すことはむずかしい．

4.2　3 階層データ保護システム

障害発生時の大容量データの短時間での “ 復旧 ” やアーカイブデータの速やかな “ 参照 ” を目的するデータを，廉価/大容量ディスクに置くことが考えられるようになった．“ 復旧 ”や“ 参照 ” のためのデータを大容量ストレージディスクに置くというシステムでは，データ保護システムを“ ストレージ-テープ ”の 2 階層システムとしてではなく，“ Online Storage（ディスク）- Nearline Storage（バックアップ専用ディスク）- Offline Storage（テープ装置）” の 3 階層ストレージシステム（図 5）としてとらえる．

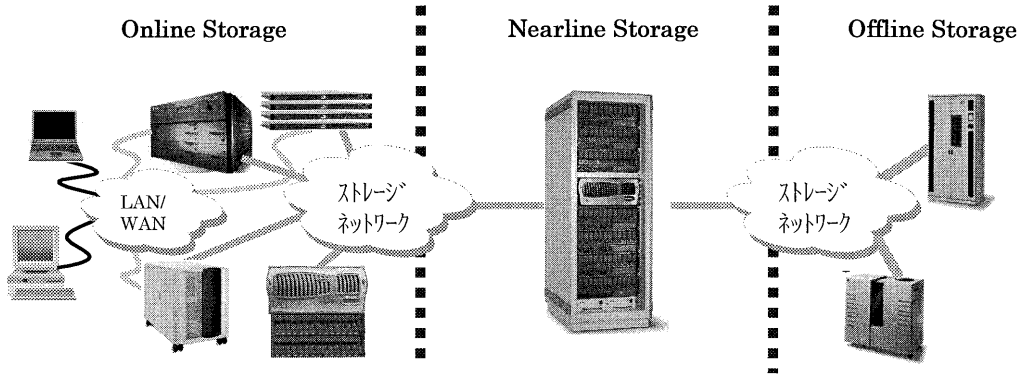

図 5　3 階層ストレージシステム

3 階層データ保護システムでは，日々の業務に頻繁に使用するデータは Online Storage に置く．その後，使用頻度が下がったデータを Nearline　Storage にステージングさせた上で，長期的には Offline　Storage に保存しておくということになる．Nearline　Storage にデータが保持される期間は比較的短く，最新の数世代だけである．このデータは，Online Storage のデータを “ 復旧 ” するためのバックアップデータであり，“ 参照 ” するためのアーカイブデータである．この後，長期保存用のデータは Offline Storage にとられる．Online Storage と Nearline Storage 間のデータ転送は，ストレージ自身の備えるリモートミラーリング機能を用いて行う．Nearline Storage と Offline Storage の間はバックアップソフトウェア製品による．　最近では，バックアップソフトウェア製品が “ ディスク-ディスク-テープ ” の経路をとるオプションの提供を始めている．

3 階層データ保護システムの形態をとることによって，Online　Storage に対してバックアップデータを非常に高速に復旧することができるようになる．また，参照のためにアーカイブデータを Nearline　Storage に置くことにより，ランダムアクセスに弱いというテープ媒体の持つ欠点が解消される．Nearline Storage 大容量ストレージ向けの製品は，昨年ぐらいから Net-

work　Appliance 社や EMC 社から出荷され始めている．これらの製品に共通しているのは，価格を下げるために ATA（AT Attachment）ドライブを使用していることである．ATA-SCSI もしくは ATA-FC ブリッジコントローラを内蔵し，ATA ディスクでありながら SCSI/FC 接続を実現している．

5. InfiniBand

5.1 PCI の限界

ネットワーク経由のデータフローの最大のボトルネックは，I/O スループットである．これを高速化しようとサーバ内部に大量のキャッシュとメモリーを備えても，バスシステムの高速化の問題につきあたった．クロック数が大幅に上がり，帯域幅が 4 倍になっても，もともとデスクトップコンピュータ向けに開発されたバスシステムである PCI（Peripheral　Component Interconnect）は，一度に二つのメンバ間のデータ転送しかできない．

さらに PCI はパラレルバスであるため，周波数が高くなると“スキューコントロール”の問題が発生する．周波数が 100 MHz を超えると，信号線と信号線の間に生じるわずかなずれ“スキュー（Skew）”が生じ，信号線の伝播遅延をコントロールすることが難しくなるのである．パラレルバスでは，複数の信号が同時に目的地に着く必要がある．周波数を大きくすると信号の波長が短くなり，各信号のずれをコントロールするのが困難になる．また，周波数を大きくするとノイズの影響も大きくなる．ノイズマージンを稼ぐためには，各信号線に対する負荷を減らす，つまり同時に送信する信号を減らすのが効果的だが，これでは共有バスの意味がなくなる．バスに代わるスイッチベースの I/O ファブリックの必要性が，何年も前から叫ばれていた．そして登場したのが，高速シリアルの InfiniBand（転送速度：2.5 Gbps，最大帯域幅：30 Gbps）である．

5.2 InfiniBand のアーキテクチャ

InfiniBand アーキテクチャは，HCA（Host　Channel　Adapter），TCA（Target　Channel Adapter），スイッチ，ルータの四つの基本装置から構成される（図 6）．HCA は，メモリーバス（従来型システムでは PCI アダプタ）を介したコンピューティングサブシステムへのインタフェースである．一方，TCA は既存の SCSI または FC インタフェースにかわる物理ハードドライブへのディスクインタフェースである．HCA と TCA は，スイッチをとおしてネットワークへリンクされる．

InfiniBand スイッチは，カスケードすることができる．スイッチはリーフスイッチなどの比較的簡単な装置かもしれないし，プロトコル変換や帯域幅集約のような複雑な操作を実行可能な独自のインテリジェンス機能を備えたディレクタースイッチかもしれない．ブリッジやルーティング機能を含む InfiniBand ファブリックのバックボーンに配置され，Fibre　Channel，SCSI，Gbit　Ethernet の各コンポーネントを統合することによって，既存装置サポートと円滑な移行パスを提供する．また，InfiniBand アーキテクチャでは，ルータによって他のネットワークへのインタフェースと従来型コンポーネントやネットワークとの間での変換が可能である．ルータは，WAN や MAN インタフェースに使用できる．

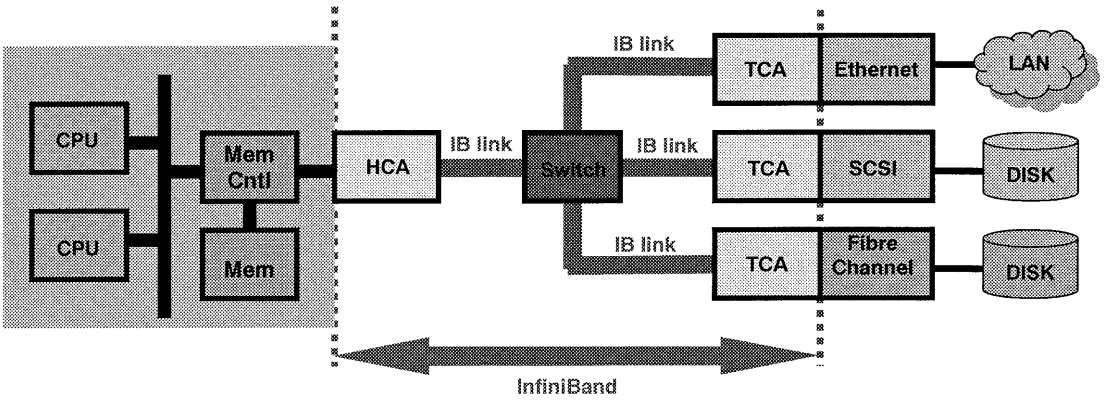

図 6　InifiniBand アーキテクチャ

5.3　InifiniBand の特徴

InfiniBand のリンク速度には，1 x の 0.5 GB 全二重リンク（一方向に 0.25 GB），4 x（2 GB 全二重），12 x（6 GB 全二重）がある．12 x のスループットは Gbit　Ethernet の理論的最大値より 50 倍高速であり，サーバでデータマーシャリングのオーバヘッドは一切かからない．

InfiniBand は複数の異なるプロトコルのパケットを流すことができ，ケーブルとして筐体の外に出すことが可能であるため，ストレージ装置やネットワーク装置といった外部 I/O 装置との接続やノード間の接続に用いることができる．そして，各種ネットワークを InfiniBand ネットワークに統合することが可能になる．従来の SAN 環境では，ネットワーク I/O，ストレージ I/O，クラスタリングシステムのノード間接続などでそれぞれのファブリックごとに異なるネットワークが存在し，異なる管理ツールを操作する必要があった．そのため，システム管理は非常に煩わしく，管理コストも膨大になっていた．さらに各ファブリックが個別のケーブルを必要とするため，サーバラックの背面は各種ケーブルが交錯し，さらに管理性を低下させていた．InifiniBand では，複数プロトコルを一つのファブリックでサポートすることが可能であり，ケーブルを大幅に削減することができる．スパゲッティのように絡み合ったケーブルを 1 本の InfiniBand ケーブルにまとめることができる．また SAN まわりのネットワークを InfiniBand で統合することによって，単一ソフトウェアで SAN の全てのノード，I/O デバイス，ネットワーク機器を管理することができるようになる．将来的には，データセンタ内を流れるプロトコルをすべてネイティブな InfiniBand プロトコルにしてしまうことによって，TCP/IP 通信によるオーバヘッドをデータセンタから排除することができる．

6.　RDMA & DAFS

6.1　RDMA

DAS や SAN のようにサーバ側にファイルシステムが存在する場合，ストレージ上のデータがサーバ上のアプリケーションに渡るまでの間に何度も繰り返しコピーされる．この間アプリケーションは自分の出した I/O 要求に対するレスポンスを待っているわけであり，このコピー操作の合計時間がアプリケーションのパフォーマンスに与える影響は決して小さいものではない．しかも 1 回の I/O イベントで転送されるデータ量はそう多くないので，10 MB ものデータの読み出しともなるとこの操作が何度も何度も繰り返し行われることになる．例えば 1

回に読み出せるデータ量が 4 KB であったとすると，10 MB のデータを読み出すには，一連のコピー操作が 2500 回も行われることになる．NAS ではストレージサブシステムにかなりのオプティマイゼーションが行われているが，何度も繰り返しコピーが行われることに変わりはない．ストレージバッファとアプリケーションバッファを直結することによりデータパスが短縮される．その結果発生するコピーの数が減り，パフォーマンスが上がる．各々の I/O 要求を一回の要求として処理し，10 MB のデータの読み出し要求を，4 KB データの 2500 回の読み出しとしてではなく，1 回の 10 MB のデータの読み出しとして処理する．これを実現するのが，RDMA（Remote Direct Memory Access）である．

InfiniBand は，TCP/IP プロトコルをサポートする．しかし，InifiniBand の高速性を最大限に活かすには，RDMA と組み合わせる必要がある．RDMA によって，異なる二つのサーバ上にあるデータベースが，OS の干渉なしに相互に通信できるようになる．VIA（Virtual Interface Architecture）は，Intel，Compaq，Microsoft によって 1996 年に提唱された高速インタフェースのためのアーキテクチャである．その特長は，OS を介さずに正確にアプリケーション間通信が行えることと，RDMA 転送が可能なことにある．VIA では，利用可能なプラットフォームの数に制限があった．VIA はサーバからサーバへデータを取得するためのプロセッサーステップ数を，数万からわずか数百へ削減した．しかし，VIA はその特有の実装方式が欠点であった．各ベンダは独自の物理ワイヤリング仕様を持っていたため，相互運用がなかった．これとは対照的に，InfiniBand では RDMA ネットワークを，イーサネットやトークンリングアダプタを使用するのと同様に使用できる．どのベンダのハードウェアが実装されるかは関係のない．アダプタは，ワイヤとプロトコルともに互換性がある．

6.2 DAFS

6.2.1 高速ファイルアクセス

RDMA の応用として興味深いのは，NAS の利点をローカル接続の raw デバイスの利点と組み合わせた DAFS（Direct Access File System）である（図 7）．DAFS は，ストレージの発展にとっては必然的な次なるステップとして，RDMA 機能のスループットを NAS の概念に統合したものである．サーバ上（および，要求処理リソース上）に存在するファイルシステムの典型的な操作は，DAFS サーバへオフロードされる．DAFS サーバ自体は，既存のインフラストラクチャ（SAN など）を使用することも，NAS 製品へ直接統合することもできる．InfiniBand ファブリックにより，データがローカルなローディスク上に存在するのと同程度の低い遅延で動作できる．

DAFS は，複数サーバで構成されたネットワーク環境で共有ファイルをローカルファイルとして，しかも高速にアクセスできることを狙っている．DAFS は，InfiniBand などの高速通信媒体の性能を最大限に発揮することを目的として新たに設計されたプロトコルであり，高速 NAS システムとして稼働する．ファイル共有においては優れていた NAS も，従来高速性の点では SAN に一歩譲っていた．しかし，DAFS を使えば，SAN を凌駕する高速 NAS システムを構築することができる．

NFS は，低帯域，高レイテンシ，コネクションレス，低信頼性の広域ネットワーク環境を扱うように設計されているため，RPC 層やソケット層といった余分なプロトコル層が存在し，特に TCP/IP の通信処理にかかるオーバヘッドは多大なものがある．それ故 InfiniBand に代

表される最新の高速通信媒体をネットワークに採用しても NFS を使用する限り，オーバヘッドの処理がネックとなり，高速通信媒体の性能を最大限に活用することはできない．

一方 DAFS は，当初から最新の高速通信媒体の性能を最大限に発揮することを目的として設計されており，高帯域，低レイテンシ，コネクション指向，高信頼性のネットワーク特性を活かした通信が可能である．通信処理にかかるオーバヘッドが最小なので，CPU の負担が少ない高スループットのデータ転送ができる．VIA の機能を活かすことにより，ローカルバッファからリモートバッファへの OS を介さないデータ転送（RDMA）やアプリケーションからトランスポート資源を直接アクセスすることが可能である．

DAFS は通信層として VIA を使用するが，OS のバイパスと RDMA 転送が可能であれば，必ずしも VI でなくてもよい．このような要件を満たす通信層を DAT（Direct Access Transfer）と呼ぶ．したがって，正確に言えば DAFS では DAT の採用により高速ネットワークファイルアクセスを実現している．

NFS ではアプリケーションのバッファからファイルシステムのキャッシュとネットワーク層のバッファへのコピーが発生している．これに対して DAFS では，ファイルシステムのシステムコール処理は OS を経由しているが，通信層が DAT であるため CPU のオーバヘッドは非常に少ない．またデータ転送には RDMA を使用するため，メモリーコピーが入らない．

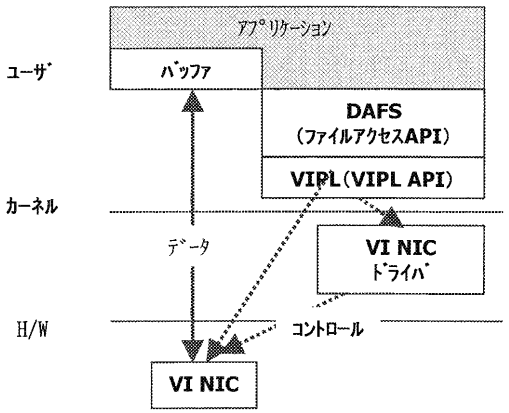

図 7　DAFS アーキテクチャ

6.2.2　DAFS の実装方式

DAFS の実装方式には，次の 3 方式がある．

fDAFS：カーネル内にファイルシステムとして DAFS を実装．

uDAFS：ユーザライブラリとして DAFS を実装．DAFS-API インタフェースを使用してファイルシステムにアクセスする．DAFS-API を使用するようにアプリケーションを変更する必要があるが，ファイル操作も OS を介さないため fDAFS よりもさらに高速である．専用システム向きである．

dDAFS：OS 中に実装した仮想的ブロック特殊ドライバが DAFS プロトコルを使用してサーバ側とのやり取りを行う．ファイルサーバ上のファイルをアプリケーションに対して Raw ディスクに見せる．DBMS など，Raw ディスクを前提としたアプリケーションを前提としている．

DAFS 要件を満たすネットワークとして cLAN，VI-IP，Myrinet などがあるが，いずれもプロプリエタリなネットワーク技術であり，高価な専用ネットワーク構築が必要である．今後最も期待できるのは，InifiniBand である．InifiniBand はオープンな規格であり，また安価である．ノード間高速通信だけでなく，I/O 装置の接続や通常の TCP/IP 通信にも使用でき，クラスタ環境内のネットワークを統一することができる．また Ethernet で RDMA をサポートする iWARP アーキテクチャの標準化が進められている．また，SCSI プロトコルを RDMA 上に流す SRP プロトコルがある．DAT 要件を満たしたネットワークで，高速なディスクアクセスが可能になる．しかしながら，ブロックレベルアクセスであり，ファイル共有はできない．ファイルサーバ/ディスクサブシステム間で使用されると思われる．

7. お わ り に

ブロードバンドネットワークアプリケーションが急速に普及し，それに伴なって新たなストレージインフラストラクチャのニーズが増大し始めている．IP ネットワーク接続ストレージは，こうしたニーズに対する回答である．本稿で紹介したさまざまな新技術を導入し統合することにより，IP ネットワークに対するこれまでの投資と知識を無駄にすることなく，TCO を徹底的に削減しながらも，拡張性，パフォーマンス，運用管理機能，可用性，セキュリティといったビジネス上の全ての利点を提供することが可能になる．

参考文献
［1］ 喜連川　優，“ストレージネットワーキング”，オーム社，2002 年 7 月 1 日
［2］ SunWorld，“特集　エッジコンピューティング”，SunWorld, Nomvem ber, 2002
［3］ SunWorld，“特集　新サーバラインアップの実力を探る　—Sun プラットフォーム 2003—”, SunWorld, April, 2003
［4］ 中村　正澄，小池　浩之，“InifiniBand のすべてがわかる！”，Network World 2002 年 10 月号，IDG ジャパン
［5］ DB 2 magazine 日本語版，“InfiniBand がスロットルを開ける”，IBM, 2002 年 2 月号
［6］ Rajesh Godbole,“Filers in a CDN Environment”, Network Appliance［TR 3104］
［7］ Iftikhar Ahmed, Cisico Systems, Inc., Rajesh Godbole and Shiva Vishwanathan, Network Appliance, Inc.,“An Open Standards Approach to Network-Centric Storage”, Network Appliance［TR 3121］
［8］ Network Appliance, Spectra Logic,“Spectra Logic and Network Appliance : A Three-Tiered Storage Architecture”, Network Appliance［TR 3153］

執筆者紹介 根　来　　元（Gen Negoro）
1982 年関西学院大学文学部心理学科卒業．同年日本ユニシス(株)入社．A シリーズ金融オンラインシステム System-F の開発・導入を担当後，A/NX シリーズの各種ランゲージプロセッサやスクリーンデザインユーティリティなどの開発・受入れ保守を行なう．ブロードバンドビジネス事業部を経て，現在は商品企画部で NAS を中心とするストレージ製品のマーケティングに従事．