■8 群（情報入出力・記録装置と電源）－ 2 編（情報ストレージ）

5 章 ストレージサブシステム

（執筆者：山本　彰）［2011 年 1 月 受領］

■概要■

ストレージサブシステム（Storage Subsystem）は，情報システムにおいて，情報を格納する役割を果たす主なシステムであり，例えば，銀行などの口座情報，鉄道，航空会社の座席予約情報などが格納される．システムは，情報を格納する千台規模のハードディスクドライブ（以下ディスクと略す），読み書き要求を処理する数十台～数百台のプロセッサ，ディスクキャッシュ，サーバとのインタフェース部，ディスクとのインタフェース部とこれらを結合する機構などにより構成される．

ストレージサブシステムは，物理的な読み書き単位であるブロックレベルのストレージシステムが主流であったが，論理的な格納単位であるファイルレベルのストレージシステムも普及している．前者は，近年，SAN（Storage Area Network）と呼ばれるストレージ専用のネットワークでサーバに接続される形態が普及している．後者は，一般的な IP（Internet Protocol）でサーバに接続されるが，代表的なシステムは NAS（Network Attached Storage）と呼ばれる．一方，企業の情報部門では膨大な情報の管理コストの増大が課題となり，これらのストレージシステムを効率良く管理するストレージ管理ソフトも一般的になってきた．

【本章の構成】

本章では，概要（5-1 節），SAN とブロックレベルのストレージシステム（5-2 節），NAS を代表とするファイルレベルのストレージシステム，ストレージ管理（5-3 節），新しい技術潮流（5-4 節），について述べる．

■8 群－2 編－5 章

5-1 SAN とブロックレベルのストレージシステム

（執筆者：松並直人）［2009 年 3 月 受領］

5-1-1 ブロックレベルのストレージシステム

ブロックレベルのストレージシステム（以下ストレージシステムと略記）とは，ホスト計算機（以下ホストと略記）が，ハードディスク（以下 HDD と略記）などから構成される記憶領域であるボリュームに対して，ブロックと呼ぶ単位で読み書きを行うストレージシステムである（**図 5･1**）．ストレージシステムは，情報システムの高性能化・高信頼化のニーズに対応するため，1980 年代後半から，RAID（Redundant Array of Independent（または Inexpensive）Disks）技術の採用による高性能化･高信頼化と，コントローラによる高性能化を進めてきた．

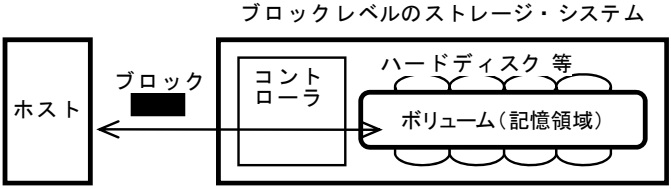

図 5・1　ブロックレベルのストレージシステムの概要

(1)　RAID 技術の採用による高性能化・高信頼化

RAID は，複数台の HDD を用いて，性能と信頼性を向上させたストレージシステムを構成する技術である．1988 年に米国カリフォルニア大学バークレイ校の David A. Patterson 教授らが提唱したのが始まりとされ，RAID レベルと呼ばれる HDD の組合せ方が考案された [1]．以下，現在広く実用化されている RAID 1，RAID 4，RAID 5 について説明する．

(a) **RAID 1**（ミラーリング）：2 台の HDD にブロック単位に複製を持ち合う．信頼性とリード性能が向上するが，2 倍の容量が必要となる（**図 5･2**(a)）．

(b) **RAID 4**：複数台 HDD にストライプと呼ぶ複数ブロックセット単位でデータを分散して格納する．更に，分散格納する複数ストライプからパリティと呼ぶ冗長データを作成し，別の 1 台の HDD に格納する．1 台の HDD に障害が発生した際にはパリティと他 HDD のデータから障害 HDD のデータを再現できる．信頼性とリード性能は向上するが，ライト時は単一 HDD へのパリティ書込みが性能ボトルネックとなる（図 5･2(b)）．

(c) **RAID 5**：RAID 4 の改善型であり，パリティも複数台 HDD に分散して格納する．パリティ格納の並列処理が可能となり，RAID 4 に比べライト性能が向上する（図 5･2(c)）．

更に近年，低価格大容量 HDD の登場により，HDD 障害時の復旧時間が長期化したことから，その期間の信頼性確保のため，RAID 6 と呼ばれる RAID レベルも実用化されている [2]．

(d) **RAID 6**：分散格納する複数ストライプから直行する 2 種類のパリティを生成することで 2 台までの HDD 障害に対する耐データ喪失性を備え，RAID 5 に比べ更に信頼性が向上する（図 5･2(d)）．

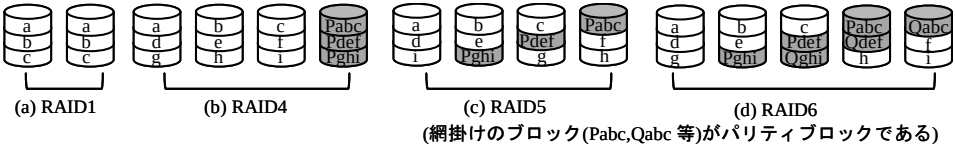

図 5・2　RAID 技術の概要

(2)　コントローラによる高性能化

上記のような RAID 技術を用いたストレージシステムは，コントローラ内にプロセッサと大容量メモリを備える．プロセッサは，RAID 技術で構成されたボリュームに対するライト・リード時に，ブロック分散／集合制御を行う．更に，プロセッサは，多くの場合，専用ハードウェアと協調して大容量メモリ上でパリティの生成や照合を行う．更に，大容量メモリは，ボリュームに格納されたデータのキャッシュとしても動作し，ホストへの高速な応答を可能とする．

5-1-2 SAN（Storage Area Network）の登場

(1)　DAS と SAN

1990 年代前半までのストレージシステムは，**図 5･3**(a)のように，ホストと一対一で接続される形態が一般的であった．このような接続形態を DAS（Direct Attached Storage）と呼ぶ．DAS で用いられるインタフェースとしては，オープンシステムに用いられる SCSI（Small Computer System Interface）[3] が一般的である．

しかし，情報システムの進展にともない，システムに格納されるデータ量は増大してきた．増大するデータをホストごとに接続されたストレージシステムで個別に格納して運用することにより，ストレージシステムでは，利用効率低減や管理負担増大という課題が生じた．この課題を解決するため，1990 年代後半から，図 5･3(b)のような，ストレージシステムとホストをネットワーク接続する SAN（Storage Area Network）という接続形態が登場した．SAN により，ストレージシステムは複数のホストで共有され，ストレージシステムの集約（ストレージコンソリデーション）による容量利用効率の向上と，管理負担の軽減が実現された．

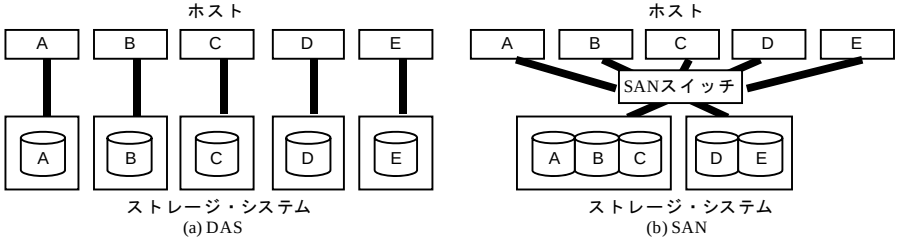

図 5・3　DAS と SAN の概要

(2) SAN を支えるネットワーク技術

SAN を構築するためのネットワーク技術としては，高速・大容量のデータ通信に耐えるものが求められる. 現在最も多く使われるのが FibreChannel である [4]. FibreChannel は ANSI T11 委員会で標準化されたネットワーク技術であり, FibreChannel 上で SCSI プロトコルや FICON プロトコル（FIber CONnection；メインフレーム用インタフェースである ESCON（Enterprise System Connection）互換のインタフェースプロトコル）を用いたデータの読み書きが可能である．FibreChannel による SAN を特に FC-SAN と呼ぶ.

FC-SAN は高速で大容量帯域を備えることが特徴であるが，SAN を構成するためのスイッチや光ファイバケーブルなどのインフラが必要であり, 相応のコストを要する. これに対し，普及が進んだ IP（Internet Protocol）ネットワークで SAN を構築する IP-SAN が考えられた．IP-SAN では，TCP（Transmission Control Protocol）の上で SCSI インタフェースを用いてデータを読み書き可能とする，iSCSI（Internet Small Computer System Interface）が用いられることが多い [5]．iSCSI による IP-SAN は，再送によるストリームデータ転送制御を基本とする TCP を用いるため，現在は性能面で FC-SAN と比べて劣るが，普及が進み低価格化した IP ネットワークインフラを利用できるメリットがある．更に，FibreChannel のように接続距離に制約がないことから，低コストによる遠隔地バックアップへの活用も期待されている．

5-1-3 ストレージシステムの機能の進化

情報システムが企業活動や社会活動のインフラとして普及することで，1990 年代頃から，システム障害だけでなく地震などの災害による情報システムの停止が，企業や社会に与える影響が甚大となり，可用性の更なる向上が求められるようになった．また，インターネットの普及やリッチコンテンツ（映像など）のディジタル化が進むことでデータ量が爆発的に増大し，大量データを更に効率的に運用することが求められるようになった．このような課題に応えるため，ストレージシステムでは，(1) コピー機能による可用性の向上，(2) 仮想化機能による運用効率向上が進められてきた.

(1) コピー機能による可用性の向上

コピー機能は，ストレージシステム自身がデータの複製（レプリカ）を生成する機能である．コピー機能は，ストレージシステム内にレプリカを生成する (a) ローカルコピー機能と，複数台のストレージシステム間でレプリカを生成する (b) リモートコピー機能に大別される.

(a) ローカルコピー機能

5-1-1 項で述べたとおり，ストレージシステム内の HDD の障害に対しては，RAID 技術によりデータを保護することができる．ローカルコピー機能は，情報システム障害により発生したデータ不整合からデータを保護するために有効である．ローカルコピー機能には，レプリカの生成方法として，ボリュームのフルコピーを保持するミラースプリット型と，ボリュームの差分コピーを保持するスナップショット型が存在する．これらは，レプリカデータを保持するための記憶容量（コスト）と性能においてトレードオフの関係にあり，データ特性に基づき，どちらの方法で運用するかが選択される.

(b) リモートコピー機能

リモートコピー機能は，あるストレージシステムが，ホスト介在なくほかのストレージシ

ステムにデータをコピーしてレプリカを生成する機能である．情報システムにおいて，通常稼働するサイトに災害発生したとき，遠隔地の待機サイトで運用を引き継ぐ際に即座にレプリカを活用できる点で有効である．

リモートコピー機能には，以下の二つの方法がある．第一の方法は，ホストからのライト I/O を受領すると，遠隔地のストレージシステムへデータのコピーが完了してから，ライト I/O 完了をホストに通知する同期型である．第二の方法は，遠隔地のストレージシステムへデータコピーが完了する前に，ライト I/O 完了をホストに通知する非同期型である．これらは，遠隔地へのデータの書込みが保証できるか否かの可用性の高さと，ホストへの応答性能の高さのトレードオフの関係にあり，保護対象のデータ特性に基づきどちらの方法で運用するかが選択される．

(2) 仮想化機能

仮想化機能は，ストレージシステムの容量効率向上と管理負担の軽減を目的とした機能である．大別すると，(a)ボリューム仮想化機能，(b)容量仮想化機能がある．

(a) ボリューム仮想化機能

データ量の爆発的な増加により，ストレージコンソリデーションした SAN システムも大規模化し，SAN に接続されるストレージシステム台数も増えるようになった．

ボリューム仮想化機能では，ボリューム仮想化機能を備える装置（仮想化装置）が，ホストと複数台のストレージシステムとの間に介在して，ストレージシステムの記憶容量を一元管理する（**図 5･4**(a)）．仮想化装置としては，(i)専用装置，(ii)仮想化機能を備える SAN スイッチ，(iii)仮想化機能を備えるストレージシステム，の 3 種類の形態が実用化されている．更に，仮想化装置はホストには複数ストレージシステムの未使用記憶領域から構成される仮想的なボリューム（仮想ボリューム）を提供する．このようなストレージシステムの機種の違いを意識する必要のない統一的な容量管理により，運用効率を向上させることを狙いとする．

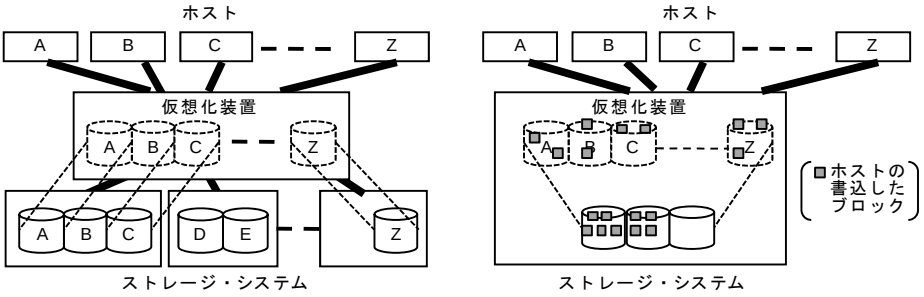

図 5・4　仮想化機能

また，ボリューム仮想化機能を用いることで，ホストからの仮想ボリュームへのアクセスを継続したまま，ブロックの格納先のストレージシステムやボリュームを変更することができる．例えば，アクセス頻度などに基づき，高価な高速 HDD ボリュームから，安価な大容

量 HDD ボリュームにデータを移行し，データ保持コストを削減する運用ができる．

(b) 容量仮想化機能

従来のストレージシステムでは，情報システムで将来使用する容量を予測し，予測に見合った容量を搭載して運用を開始していた．しかし，インターネット普及などを背景にデータ量がダイナミックに急増するようになり，安全をみて十分な容量を予測調達することは導入コストを大幅に増大させた．

容量仮想化機能とは，容量仮想化機能を備えるストレージシステムが，ホストの書込みのあったブロックだけで構成される仮想ボリュームを提供する機能である（図 5･4(b)）．ストレージシステムは，ホストからの仮想ボリュームへのデータ書込みに応じて，仮想ボリュームの書込み先ブロックに実際の HDD 上のブロックを割り当てる．本機能により，仮想ボリュームがホストに提供する記憶容量よりも小さな記憶容量をストレージシステムに搭載して運用を開始し必要に応じて拡張することが可能となった．

また，大容量のストレージシステムと大規模 SAN の安定稼働ならびに効率的運用管理を支えるためにストレージ運用管理機能が存在する．ストレージ運用管理には，障害管理，構成管理，性能管理やバックアップ運用管理などがあり，運用管理方法の標準化（参考文献 6）参照）なども進められている．また，SAN だけでなく後述するファイルレベルストレージである NAS（Network Attached Storage）も一元的に管理する SAN/NAS 統合運用管理などもあるが，詳しくは次節以降を参照されたい．

■参考文献

1) D. Patterson, et al., “A Case for Redundant Arrays of Inexpensive Disks(RAID),” ACM SIGMOD conference proceedings, Chicago, IL, pp.109-116, 1988.
2) Peter M. Chen, et al., “RAID: High-Performance, Reliable Secondary Storage,” ACM Computing Surveys, vol.26, no.2, pp.145-185, 1994.
3) インターフェース編集部(編), “SCSI 完璧リファレンス,” OPEN DESIGN, no.1, CQ 出版社, 1994.
4) JSDF ファイバチャネル技術部会, “ファイバチャネル技術解説書 II,” 論創社, Oct. 2003.
5) J. Satran, et al., “Internet Small Computer Systems Interface (iSCSI),” IETF RFC 3720, 2004.
6) ISO/IEC 24775:2007, “Information Technology - Storage management,” 2007.

■8 群－2 編－5 章

5-2 NAS を代表としたファイルレベルのストレージシステム

（執筆者：岩嵜正明）［2009 年 8 月 受領］

5-2-1 NAS の概要

NAS は Network Attached Storage の略称であり，IP ネットワークを介して，NFS（Network File System）や CIFS（Common Internet File System）などのファイル共有プロトコルによって，ホストコンピュータ（クライアント）と接続されるストレージを指す．NAS はファイル共有サービスに最適化された専用サーバマシンと捉えることもできる．

NASはクライアントとの接続に標準化されたファイル共有プロトコルを採用しているため，ブロック入出力型ストレージと比較すると，複数のクライアント間でのデータ共有を容易に実現できることが特長である．また，クライアントとの接続にファイバチャネル用HBA（ホストバスアダプタ）やファイバチャネルスイッチを必要とせず，広く普及しているEthernet[*1]コントローラやEthernetスイッチを利用できる点も，コスト面で有利といえる．

現在，NAS の範疇には，ミッションクリティカルな業務に適用できるハイエンドから，個人や家庭向けのローエンドまで，幅広い製品群が存在する．後者は，家庭向け無線ルータと一体化された NAS，あるいは，HDD（Hard Disk Drive）ビデオ録画装置と一体化された NAS を始めとして，単純な NAS の枠組みを超える高機能機器も登場している．また，多くの NAS がファイルアクセスプロトコルとして，NFS や CIFS に加え，FTP（File Transfer Protocol），HTTP（HyperText Transfer Protocol），WebDAV（Web-based Distributed Authoring and Versioning）などのプロトコルを提供している．

5-2-2 NAS の高信頼化

NAS はユーザのデータを格納する装置であるため，ローエンドの一部を除く大部分の製品は，高信頼化機能を提供している．具体的には，ほとんどの NAS が RAID（Redundant Arrays of Inexpensive Disks）構成ストレージを採用して HDD のハードウェア障害に対処している．更に，ミッションクリティカル用途へ適用する NAS では，2 台の NAS ノードをペア構成とする HA クラスタ（High Availability Cluster）機能を提供し，ノード間でフェイルオーバ動作を可能とするのが一般的である．また，クライアントと接続する Ethernet リンクの冗長化機能を提供する NAS も多い．

二つのノード間で相互にフェイルオーバ可能とするためには，障害発生時に相手ノードのファイルシステムを引き継いで動作できる必要がある．これを実現する手法としては，**図 5･5** に示すように二つのノード間でデータの書込みを常時ミラーリングするか，あるいは，**図 5･6** に示すようにコントローラ部分と RAID 構成ストレージ部分を切り離して両者間をファイバチャネルなどで接続する必要がある．後者の場合，コントローラ部分と RAID 構成ストレージ部分を接続するファイバチャネルパスの冗長化機能を提供する NAS が多い．

上記の高信頼化技術に加え，NAS の多くはジャーナル方式のファイルシステムを採用しており，予期しない停電やハードウェア障害が発生しても，ファイルシステムの論理的整合性が破壊しないように配慮している．

*[1] Ethernet は富士ゼロックス社の登録商標です．

なお，多数の NAS ノードを接続し，これらのノード間でファイルをストライピングして冗長性をもたせて分散格納する RAIN（Redundant Arrays of Independent Nodes）方式の NAS システムも登場している．

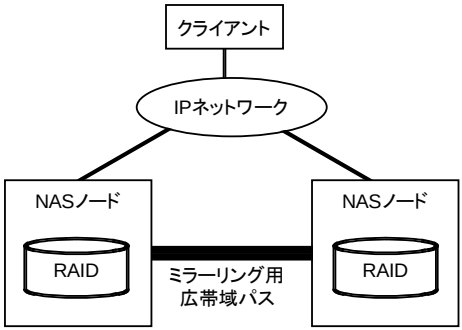

図 5・5　ミラーリング方式の HA クラスタ構成

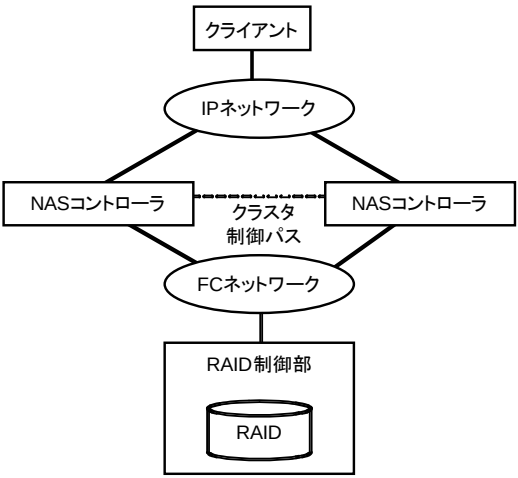

図 5・6　外部ストレージ共有方式の HA クラスタ構成

5-2-3 NAS の高性能化

NAS の多くは，ファイルサーバとして運用される汎用サーバに対抗する優位点として，ファイル共有サービスに特化した性能最適化を施している．具体的には，HDD アクセスの遅延時間の影響を低減すべく，一般的な汎用サーバに比較すると大量のバッファメモリ（半導体メモリ，DRAM）を搭載している．更に，バッファメモリと HDD のアクセス速度差を吸収すべく，両者の中間層にフラッシュメモリなどの不揮発性メモリを採用している NAS も多い．

なお，ファイル共有サービスに特化したハードウェアアクセラレータを採用したり，ソフトウェアの高度な最適化を施して高性能化を図った NAS も存在する．しかし，CPU を始めとする汎用サーバ構成部品や汎用 OS の性能が著しく向上し，これらの NAS の性能面での優位性は縮小しつつある．

5-2-4 NAS のデータ保護機能

多くの NAS は，ブロック入出力型ストレージと同様，差分スナップショット機能やリモートコピー機能といったデータ保護機能を提供している．

NAS における差分スナップショットの最も単純な実装は，更新が発生したファイルをファイル単位で差分ボリュームにバックアップする方式である．しかし，この方式では，データベーステーブルに代表される巨大ファイルに対してランダムな書込みが発生する用途では，スナップショットの保持に必要な差分ボリュームの容量が増大してしまう．このため，多くの NAS では，ファイル内の更新が発生したブロックだけを選択的に保持することにより，差分ボリュームの容量を低減させる方式を採用している．

ブロック入出力型ストレージと比較して，NAS のスナップショット機能の優位性はそのリストア機能にある．ブロック入出力型ストレージと異なり，NAS のリストア機能ではファイル単位やディレクトリ単位にリストア対象をきめ細かく指定できる．すなわち，ファイルシステム全体（ここではボリュームと同義）をある時点のスナップショットの状態にリストアするだけでなく，特定のファイルやディレクトリだけを選択的に指定した時点の状態にリストアすることができる．この柔軟性により，ハード障害によるファイルシステム全体のリストアといった目的に加え，「エンドユーザが誤操作によって必要なファイルを削除してしまった場合に，当該ファイルだけをリストアしたい」といった要求に容易に対応できる．

リモートコピー機能に関しては，元来，NAS はファイル共有サービスの基盤として TCP/IP ベースの通信機能を備えており，FC-SAN（Fibre Channel-Storage Area Network）ベースのブロック入出力型ストレージに比較するとファイバチャネルエクステンダ（FC-SAN と IP ネットワーク間のプロトコル変換を行う装置）などが不要であり，容易かつ低コストに遠隔バックアップシステムを実現できる．

5-2-5 NAS の容量管理機能

多人数でのファイル共有に利用される NAS では，ディレクトリごと，あるいは，ユーザごとに容量制限を設定できるクォータ管理機能を提供している．クォータ管理機能を利用すると，予め管理者が設定した上限容量を超えるクライアントからの書込み要求を拒否でき，NAS システム全体が容量不足に陥ることを回避できる．

また，一部の NAS は容量仮想化機能（動的に容量割当てを拡張する機能）を提供している．一般に，NAS は運用開始時に各ファイルシステムの容量を決定する容量設計が必要である．しかしながら，ファイルシステムの必要容量を正確に予測することは必ずしも容易ではない．導入時の容量設計が不適切で，容量が不足するファイルシステムが生じたり，逆に，余剰な容量を使い切れないファイルシステムが生じるという問題が発生しやすい．この問題を解決するために，容量仮想化機能では，物理容量をプール化し，各ファイルシステムには仮想的な容量を割り付けておき，実際に使用された容量分だけの物理容量を割り当て，容量効率を

改善する．ただし，容量仮想化機能を利用して，NAS システムの物理的な容量増設を繰り返すと，物理容量と仮想容量の対応が複雑化してしまう問題がある．

上記の容量管理機能に加え，近年の急速なファイルシステム容量の増大にともない，データ圧縮機能やデータ重複排除機能を備えた NAS も登場している．

5-2-6 NAS のアクセス権限管理機能

近年，情報セキュリティの重要性が認識されるにともない，NAS のアクセス権限管理機能も強化されている．例えば，従来の標準的なファイル所有者・グループ・第三者によるパーミッション指定と比較して，より柔軟でセキュアなアクセス制御を可能とする POSIX ACL（Access Control List）機能などを提供する NAS が多い．また，ACL 機能を利用するには，NAS クライアントも含めたユーザ認証システムとの連携が不可欠であり，NAS の多くが LDAP（Lightweight Directory Access Protocol）などの外部認証サーバとの連携機能を備えている．

また，企業活動に対する法的規制の強化に対応し，一部の NAS では WORM（Write Once Read Many）アーカイブ機能を提供している．これは，NAS 内に格納されたファイル群を，設定した期間にわたって，いかなる管理者でも書換えや消去ができない状態にする機能である．

5-2-7 NAS システムの機能

NAS 単体の機能ではないが，複数の NAS を組み合わせたシステムにおいて，下記に示す機能がある．

- **オンラインマイグレーション機能**：NAS システムの導入時や増設時には，大量のファイル群を NAS 間で移動する必要が生じる．一部の NAS システムは，ファイル共有サービスを停止させずに，ファイル群を移動できるオンラインマイグレーション機能を提供している．
- **階層制御機能**：一部の NAS システムは，ビット単価の異なる複数の NAS を組み合わせて階層ストレージを構築し，アクセスされなくなったファイルをビット単価の低い階層へ自動的にアーカイブする機能を提供している．
- **GNS**（Global Name Space）**機能**：一部の NAS システムは，複数の NAS を並置し，各 NAS のファイルシステムを統合して，仮想的に巨大な単一ファイルシステム空間を提供する GNS 機能を提供している．GNS 機能によって，性能や最大容量のスケーラブルな拡張が可能となる．

■参考文献

1) RFC0959 File Transfer Protocol.
2) RFC1813 NFS Version 3 Protocol Specification.
3) RFC3530 Network File System (NFS) version 4 Protocol.
4) RFC2616 Hypertext Transfer Protocol -- HTTP/1.1.
5) RFC4918 HTTP Extensions for Web Distributed Authoring and Versioning (WebDAV).
6) RFC4511 Lightweight Directory Access Protocol (LDAP): The Protocol.

■8 群－2 編－5 章

5-3 ストレージ管理

（執筆者：吉田　浩）［2009 年 3 月 受領］

5-3-1 ストレージ管理の要件と対象範囲

SAN，NAS などのストレージネットワーキングの普及が進むにつれ，ストレージ管理の重要性が増大している．当初，ストレージネットワーキングは，ストレージの統合化・集約化によって運用管理コストの低減に貢献した．しかし，ストレージネットワークに接続されるストレージ装置数やサーバ数の増加，ストレージ装置の大容量化・高機能化，格納されるデータのビジネス上の重要性の増大によって運用管理コストが上昇し，結果として，利用者はハードウェアの価格性能比の向上（ストレージにおいては容量単価の低減）を十分享受できないばかりか，ビジネス上の価値を生み出すストレージ投資が困難となる状況が生じている．このため，ストレージ管理の省力化・運用管理負担の低減が強く求められるようになっている．

ストレージの管理機能は，以下の二つに大別される．

(1) データを格納する「入れ物」あるいは「道具立て」の管理．すなわち，ストレージ装置（ディスクアレイ，テープライブラリ，スイッチなどのストレージネットワーク装置，ホストバスアダプタなど）と，それらを使う際に必要となるアクセスパス，論理ボリュームなどの管理機能と，管理を支援する機能．具体的には，ストレージ機器・構成管理，性能管理，障害管理，あるいはストレージの使用状況を管理するストレージ資源管理など．

(2) ストレージに格納されるデータの「中身の管理」．「データ管理」と呼ばれることもある．代表例は，データのバックアップ／リカバリーなどの「データ保護」．

実際には，例えばストレージ資源管理では格納されているデータの種別を意識するなど，データの中身に多少踏み込んだ処理が必要な場合もあり，この意味で，両方のカテゴリにまたがる管理機能・製品もある．なお，狭い意味でのストレージ管理としては(1)を指すことも多く，本節でもこのカテゴリの機能を中心に説明する．

5-3-2 ストレージ装置・構成管理，性能管理，障害管理

ストレージ装置及びストレージネットワークにおけるハードウェアを中心とした管理機能である．ストレージネットキングの進展にともない，ストレージ装置そのものの管理だけではなく，スイッチ・ネットワーク構成・トポロジーの管理など，ネットワークとしての管理機能が重要となってきた．しかも，ネットワーク特有の問題として，性能問題，障害の根本原因，その影響範囲などを把握するのが難しいということがあり，このような観点から性能管理や障害管理の重要性が増している．

以下に機能分類の一つの考え方を示すが，実際のストレージ管理ソフトウェア製品では，ここにあげた複数の機能を適切に組み合わせ，運用管理コストの低減を図っている．

(1)　機器管理・構成管理

ストレージ装置（ディスクアレイ，テープライブラリなど），ネットワーク装置（FC ス

イッチなど），サーバ（ホストバスアダプタなど）などの SAN の構成要素に対して各種の設定と状態監視・表示を行うとともに，構成要素間の接続関係の設定・監視・表示を行う．例えば，ディスクアレイ装置管理機能としては，装置状態監視・表示，LUN の作成・設定，LUN マスキング・LUN マッピングの設定などが提供される．スイッチ管理や SAN 構成管理としては，スイッチのポートの状態監視・表示，ポートの統計情報の取得・表示，ファブリックのトポロジー検出・表示，ゾーン設定，複数スイッチの一括管理などの機能が提供される．

(2) 性能管理

ディスクアレイや FC スイッチが取得する性能統計情報などを元に，SAN を通じたストレージアクセスの性能を監視し，表示・分析する．装置の性能パラメータ（ビジー率など）が既定のしきい値を越えた場合に警報を発するといった監視機能も含まれる．

(3) 障害管理

ディスクアレイや FC スイッチの状態を監視し，ハードウェアの故障などの異常が発生した場合の表示や，管理者や保守センタへの通知などを行う．

ストレージ管理に限らず，運用管理全体の方向として，管理対象となる IT 資源と，その資源の上で実現されるビジネス，業務とを関連付けることが重要となってきており，関係管理や可視化といった技術が重要となっている．例えば，障害管理機能では，ある装置や部品の故障が，どのアプリケーションにどう影響するかを分析・表示して，業務のサービスレベルの維持に役立てるといった機能強化が行われている．更に，性能管理とデータベース管理システムのチューニング機能とを組み合わせて，データベースなど特定のアプリケーションのアクセス性能を分析し，チューニングに役立てるといった機能もある．

5-3-3 ストレージ資源管理

ストレージ資源管理（Storage Resource Management：SRM）とは，SAN 上のストレージ装置（ストレージ資源）の使用状況を把握・管理する機能である．

SAN に複数のストレージ装置が接続されている場合，しばしば装置ごとの使用量が不均衡になりやすい．すなわち，ある装置はほとんど空き容量がないのに，別の装置はかなり空き容量があるといった状況が起きる．また，同じデータが重複して格納されていたり，ビジネス上不要になったデータが残っているといった状況も多い．このような使用状況を的確に把握しておかないと，本来ならば現状で賄えるのに新規にストレージ装置を増設してシステム全体のストレージ資源の使用効率が低下する，あるいは，ストレージ資源の増設が適切なタイミングで行われず業務がスペース不足で停止してしまう，といった問題が生じる．

ストレージ資源管理は，SAN 上の全ストレージ資源の検索と使用状況調査，それらの統合監視・総合操作，ストレージ使用量の表示・レポート・分析，使用傾向の分析・容量増加予測，それに基づく警告や自動化スクリプトの実施などの機能をもつ．更に，ストレージ資源管理によるアプリケーションごとの使用状況の把握と予測に基づいて，アプリケーションにストレージ資源を追加割当てする操作や，ストレージ装置の再編成を行う操作などを自動化する機能があり，プロビジョニングと呼ばれる．これは，アプリケーションが要求するサービスレベルを維持するように設定されたポリシーに従って，アプリケーションとストレージ

資源との関連付け，ストレージの自動設定・構成などを制御するものである．

5-3-4 データの保護

広い意味でのストレージ管理はデータ管理を含むことを初めに述べたが，ここでは，データ管理の中でも最も重要で普遍的に行われているデータの保護について述べる．

データの保護とは，物理的あるいは論理的なデータの破壊や遺失に対して，データを復元して業務を再開できるようにすることである．ハードディスクやストレージ装置の障害といった物理的なデータの破壊に対しては，ハードウェアの多重化だけで対処できる場合もある．しかし，ソフトウェアの障害によって誤ったデータが上書きされてしまうといった論理的な破壊まで想定し，更に対処に必要なコストの妥当性を考慮して，データのコピーを別の媒体・装置・場所に保管するバックアップと，それに基づいてデータを復元し業務を再開するリストア／リカバリーが，データ保護としてごく一般的に行われている．

バックアップ／リカバリーの技術は多岐にわたり，ここですべてを網羅することはできないが，管理という視点からは，以下のように整理できる．

(1) ストレージネットワーク構成と関連付けたバックアップ管理

ストレージネットワーキング環境では，業務用 LAN を経由せずに SAN を使用してデータを転送する LAN フリーバックアップ，SAN に接続された専用のバックアップサーバ経由あるいはストレージ装置間で直接バックアップを取得することで業務サーバに負荷を与えないサーバフリーバックアップなどのバックアップ技法が適用される．これらを実行するために，バックアップの視点からストレージネットワーク上のサーバ及びストレージ装置の構成を把握・管理する．

(2) バックアップの実行管理

一般に，バックアップは定期的あるいは業務処理の区切りなどのスケジューリングに従って採取することが一般的である．一方，ディスクストレージ装置では，バックアップにともなうアクセス停止時間（バックアップウィンドウ）を極力短縮するために，ストレージ装置内の特定時点のデータをそのまま保存するポイントインタイムコピーあるいはスナップショットと呼ばれる機能が一般的に提供されるようになっており，バックアップを採取する際には，このような機能の制御が必要である．更に，データベースのバックアップでは，バックアップ採取前に一旦データベース内のデータの一貫性をとり，バックアップ中（ポイントインタイムコピー採取中）はデータの変更を停止し，バックアップ採取後にアクセスを再開するといった処理手順が必要であり，データベース管理システムとの連携が行われる．このようなバックアップ処理のマクロなスケジューリングと，バックアップ処理中のデータベースとの連携やポイントインタイムのミクロな実行制御は，バックアップ管理ソフトウェアの重要な機能となる．

(3) バックアップ世代／媒体の管理

定期的あるいは業務処理の区切りでバックアップを採取することによって，一般的には複数の世代のバックアップが作られる．また，最新の世代のバックアップは，バックアップ採

取とリストアの高速性を考慮してオンラインストレージ装置のディスク上に採取するとともに（D2D バックアップ），ストレージ装置自体の障害に備えてニアラインストレージ装置やテープライブラリ装置に更にコピーを作り（D2D2T バックアップ），一方，古い世代のバックアップは低速・安価なオフラインストレージに格納するといったように，世代によって，コピー数，格納装置，媒体を変えたり，災害対策として遠隔地にバックアップを置くといったこともよく行われる．このように，あるデータの単位に対して，どのようなバックアップ世代がどこにどのような形態で存在しているかを管理することも必要となる．

5-3-5 ストレージ管理における標準化

ストレージがもっぱら DAS（Direct Attached Storage）であり，しかもサーバとストレージ装置が同一のベンダから供給されていた時代（メインフレームによって代表される）には，そのベンダ専用のハードウェアの添付品として提供される DAS 管理ソフトウェアを使って，ストレージ管理を行うのが一般的であった．しかし，ストレージネットワークは，いろいろなベンダのストレージ装置やネットワーク装置によって構成されるのが一般的であり，ストレージ管理ソフトウェアには，マルチベンダ環境への対応が求められる．

マルチベンダによって構成されるストレージネットワークの管理では，ストレージ装置とストレージ管理ソフトウェア間のインタフェースに関する標準化が必要となる．管理ソフトウェアと管理対象間の情報取得や操作で多用される標準プロトコルの一つに SNMP（Simple Network Management Protocol）がある．SNMP によってプロトコル自体は共通化されるが，ここで情報を授受するための MIB（Management Information Base）のデータ構造がベンダによってまちまちであり，これだけでは，マルチベンダストレージ管理は十分には実現できない．このことは，単にプロトコルやインタフェースの標準化だけではなく，これを通じて授受される情報や操作の意味，すなわち管理対象のオブジェクトモデルを標準化する必要がある．

このような理由から，ストレージの普及促進・標準化団体 SNIA（Storage Networking Industry Association）は，2002 年から，マルチベンダのストレージ管理仕様 SMI-S（Storage Management Initiative Specification）の標準化を進めてきた．SMI-S は，2008 年 10 月現在では，1.3 版が完成しており，ISO の標準としても採用されている．

SMI-S では，運用管理関連技術の標準化団体 DMTF（Distributed Management Task Force）が制定し運用管理で多用される CIM（Common Information Model）に基づいて，ディスクアレイ，スイッチなどの管理対象種別ごとに共通のオブジェクトモデル（プロファイル）を規定している．プロファイルでは，装置種ごとに，標準的な特性情報とその形式，標準的にサポートされる操作とそれによって実現される動作を規定している．例として，「アレイ」のプロファイルでは，装置や搭載ディスクに関する取得可能な情報と，ディスクの割当て，LUN の作成や LUN マスキング／LUN マッピングなどの機能が規定されている．また，「ファブリック」のプロファイルでは，装置の検出（ディスカバリ）とトポロジー情報取得，ゾーンの検出・構成・制御などの機能が規定されている．更に，管理対象と管理ソフトウェア間の情報や操作の授受を，やはり DMTF が規定する Web ブラウザベースの運用管理基盤 WBEM（WeB-based Enterprise Management）によって行うものである．

SMI-S は，主要なストレージベンダのほとんどが製品に実装しており，ストレージ管理の

標準として確立しつつある.

■参考文献

1) 喜連川　優, 他,“ストレージネットワーキング技術,”オーム社, 2003.
2) “Storage Management Technical Version 1.3.0,” Storage Network Industry Association, 2008.
3) JDSF データバックアップソリューション部会, “SE のためのバックアップ&リストア,” IDG ジャパン, 2002.

■8 群－2 編－5 章

5-4 ストレージ技術の新潮流

（執筆者：大枝　高）［2009 年 9 月 受領］

5-4-1 オンラインストレージサービスとそれを支えるプラットフォーム

従来から存在したネットワークコンピューティング，ユーティリティコンピューティングや XaaS などが，安価なブロードバンドなどの普及により発展したクラウドコンピューティングが注目されている．IT の所有から利用への流れがクラウドコンピューティングにより加速し，私的なデータをインターネットの先にあるサービスに託すことに対する抵抗感がなくなりつつある．Amazon の S3[1] などに代表されるオンラインストレージサービスはこのような背景のもとで登場した．

オンラインストレージサービスの特徴の一つはその低価格性で無料のサービスも多数存在する．広告による収入など，ビジネスモデルの工夫による低価格化という側面もあるが，オンラインストレージを支えるインフラストラクチャの低コスト化がなければ，無料もしくはそれに近いサービスや使用量に応じて課金するサービスは実現できない．2000 年前後に一時話題になった SSP（Storage Service Provider）でも使用量課金のサービスを提供したが，ストレージ装置を所有するのに比べた際のコストメリットが十分でなかったため普及しなかった．

オンラインストレージサービスの低コストを支えるストレージはどのようなものであろうか．比較するため，1990 年代後半から Fibre Channel の登場とともに普及した SAN（Storage Area Network）環境でのストレージの特徴を整理すると，(1) 異種のサーバから専用ネットワーク（SAN）で共有，(2) 大規模ストレージにデータを集約し，容量管理，バックアップ，災害対策用リモートレプリケーションなどの運用を一元化，などがあげられる．これらの特徴により当時ストレージ装置価格の数倍に上っていたストレージ管理のコストを抑えることができ，大規模ユーザから急速に普及が進んだ．

これに対し，Amazon，Google やオープンソースで提供されている Hadoop[2] におけるストレージ層の特徴は，(1) コモディティサーバの内蔵ディスク，もしくは機能の低い JBOD（Just Bunch of Disks）にデータを分散して格納，(2) HDD 障害への耐性，均一なサーバ群から効率良いアクセス性を提供するファイルシステム（GFS：Google[3]，HDFS：Hadoop など），データベース（HBase：Hadoop），などがあげられ，ストレージ装置のコストを抑え，上位の分散ファイルシステムなどのミドルウェア層でストレージ管理・データ管理を実装している．

5-4-2 省電力ストレージ技術

IT の世界でも省電力技術が注目されるようになってきた．IT 機器の消費量は 2006 年で国内の総消費量の 5 %だが，2025 年には約 20 %までその割合が増大すると予測されている（出展：経済産業省，“情報通信機器の革新的省エネ技術への期待”，グリーン IT シンポジウム 2007）．国内のデータセンターの消費電力の内訳を見ると，IT 機器が 45 %で，空調 30 %，給電 18 %などとなっており，設備を含めたトータルな対策が必要なことが分かる（出展：JEITA，2008 年 5 月）．

米国でも 2005 年頃からデータセンターでの電力・空調の問題が関心を集めている．日本同様，将来，消費電力の増加に供給能力が追いつかなくなるという懸念も指摘されているが，

サーバ仮想化技術の普及とともにサーバ集約，更にデータセンターの集約を進めるのに既存のデータセンターの電力容量・冷却能力が不十分で狙いとした集約ができないという，より差し迫った課題として議論されている．

このため，直接的な省電力だけでなく IT リソースの利用効率を高める技術が省電力技術として捉えられている．その例が，容量仮想化，重複排除技術，大容量 SATA（Serial Advanced Technology Attachement）や SSD（Semiconductor Storage Device）などを利用した階層ストレージである．一方，直接的に電力を削減する機能として MAID（Massive Array of Idle Disks）[4] 技術がある．

容量仮想化はサーバの OS に対して割り当てる LU（Logical Unit）の容量を仮想化し，ストレージ装置側では容量をプール化して管理し，サーバから実際に書き込まれたデータにのみストレージ容量を割り当てる技術である．使用する容量が増大していくことが見込まれるサーバでは，通常余裕をもって LU の容量を割り当てておくことでシステム停止をともなう容量拡張の頻度を減らし，リソース不足によるサービス停止のリスクを軽減する．このため，容量の利用率は平均的に 30 %程度に止まるといわれている．容量仮想化を用いることでサーバごとに容量の余裕をとる必要がなくなり，更にサーバの構成変更をせずにストレージ装置での容量プールを増設することができるため，容量の利用率を上げることができ，結果として省電力化が実現できる．

また，高性能で注目されている SSD は，機械的可動部分をもたないため従来の HDD（Hard Disk Drive）に比べ容量当たりの消費電力という点でも優れており，性能当たりの消費電力という観点では更に優位である．SSD は容量や書換え回数の制約から音楽端末，モバイル用途 PC などから普及が始まっているが，今後，半導体技術の進歩によりエンタープライズでの使用も広がってくるであろう．データベース環境でのベンチマークで，主記憶に DB を常駐させた場合に対し，HDD だと 90 %程度性能が低下するが，SSD の場合は 5 %程度に止まるなどの報告が出てきている [5]．

バックアップやアーカイブなどアクセス頻度が少なく，高い性能を必要としない用途では大容量の SATA ディスクの利用が広がっている．2009 年の時点ですでに 3.5 インチで 2 TB の製品が出荷されており，今後も容量の伸びが期待される．これに対し，高性能エンタープライズ用途で主に使用される FC，SAS の 3.5 インチの装置の容量は 450 GB に止まっている．回転数は大容量 SATA ディスクで 7200 rpm，高性能 FC ディスク 15000 rpm などと差があり，SSD と合わせて記憶階層を構成することができる．エンタープライズでも高性能を要求しない非構造化データの割合が増大しているため，適材適所でディスク装置を使い分けることで容量当たりの消費電力，コストを適正化することができる．このための記憶階層管理技術は今後，自動化の進展が期待される．

一定期間アクセスがないことが見込まれるデータを格納する HDD の電力供給を停止することでストレージ装置の低消費電力化を図る MAID 技術は，VTL（Virtual Tape Library）や HPC（High Performance Computing）などの特定の用途から普及が進んでいる [4]．ピーク電力を低減するため RAID を構成する複数台の HDD は一般に同時に起動すること避けるため，単純な HDD の復帰時間より MAID が電源停止からスタンバイになるまでの時間は長く，SCSI コマンドを発行する OS がタイムアウトしてしまう可能性がある．VTL ならデータを利用する前にテープのマウント操作が入るため，この問題を回避することができる．HPC の分野で

は独自のジョブ管理ソフトなどと MAID の管理ソフトを連携させ，これから投入するジョブで使用するデータセットを格納する領域を予めスタンバイさせるなどの運用を構築することで MAID 技術を活用することができる．HDD 単位での電源制御から，HDD 筐体単位での電源制御に拡張することで，更に省電力効果を高めた例もある（出展：“データベースシステムの問い合わせ実行計画を利用したディスクアレイ省電力化に関する一考察”，喜連川他，DEWS 2007）．ストレージ装置の省電力状態からの復帰時間高速化や，上記の例のようなアプリケーション管理と電源制御管理の連携の進展により，MAID 技術は更に広がる可能性がある．

ストレージにおける省電力技術の標準化としては，SNIA（Storage Networking Industry Association）の Green Storage Working Group で各社のストレージの電力を実測するとともに，電力効率指標や測定手順の策定を進めている．EPA（Environmental Protection Agency，米国環境保護庁）ではストレージ省電力の基準作りを SNIA と連携して進めており，ストレージ版 Energy Star 制定に SNIA の Green Storage Working Group の成果が反映される見込みである．

先に述べたように，省電力技術の重要な目的の一つには，データセンター統合を進め，電力やスペースだけでなく IT リソースの点でも効率を上げることがある．このために，ストレージ単独でなく，ストレージを含めた IT 機器や冷却などの設備をモジュラー化して，設置場所の自由度と拡張の自由度を上げるアプローチも実現化されている [6]．

5-4-3 重複排除技術

前節でも取り上げたが，省電力技術というよりデータ量の増大に対応する有効な手段として期待が高まっているのが重複排除技術である．2000 年代後半から VTL 用途の製品や WAN 対応のバックアップ用途の製品，コンテンツアーカイブ用途の製品などから適用が始まり，条件によるが，従来の LZH などの圧縮では実現できない高い圧縮率を実現できることから注目されるようになってきた．数ワードなど比較的短いビット列の出現頻度の偏りを根拠に圧縮する従来の方式に比べ，重複排除技術ではファイル単位，ファイルを固定長や可変長などで分割した比較的長いビット列での一致を検出し，一致した片方を削除することでデータ量を圧縮する．

初期の論文や製品にはファイル単位の重複排除技術が多く，コンテンツアーカイブ用製品でコンテンツ識別のためのファイルハッシュを利用して重複を検出している．コンテンツアーカイブ用製品ではコンテンツ管理（真贋性保証など）のためにファイルハッシュを用いているものが多いため，この方法は比較的適用しやすい．ただし，ファイルに一部でも更新が入った場合は重複排除できないため，フルバックアップを何世代も保持するような環境ではファイルを更に分割した単位でも重複排除できる方式が適用されることが多い．重複排除の技術は，圧縮効率を高めるためのデータ分割及び一致検出などの方式の改善が進められている．更に高速なバックアップを低コストで実現するため，PC サーバの限られた CPU，メモリリソースでオンザフライ処理を行うための方式改善が行われており，4 コア，8 コアプロセッサの世代では現実的な搭載メモリで 400～800 MB/s の処理速度が実現できると予測されている [7]．

バックアップへの重複排除技術の適用効果は，ファイル単位での重複検出の場合，圧縮効果だけ考えれば従来からある差分バックアップと変わりはない．しかし，リストア運用の容

易さ，早さなどから一定期間ごとにフルバックアップを取得するのが通常であり，例えば直近の 1 週間は差分バックアップを保存しておき，それより過去については 1 週間ごとのフルバックアップを数世代保存するなどの運用が一般的である．この運用を変えずに，すべて差分バックアップをとった場合，並のデータ量に削減できることがファイル単位での重複排除を適用した場合のメリットである．更に，細粒度の重複排除技術を適用することで，データ量の削減効果が向上する．

5-4-4 サーバ仮想化とストレージ機能の連携

サーバの利用効率を高め，CPU リソースの柔軟な割当てを可能にするサーバ仮想化技術とストレージ機能は直接的な連携はあまりなく，それぞれ独立に開発されている．しかし，サーバ，ネットワーク，ストレージ全体で利用効率を上げ，リソースの柔軟な割当てができるようにするニーズは高く，それに関連する技術がいくつか出てきている．

FCoE は，Ethernet の物理層の上で FC のプロトコルを実装する技術である．サーバ仮想化でサーバ統合が進むことにより，ネットワークの物理層を Ethernet に集約するニーズが高まる．FCoE は，ノンブロッキングを保証するフレーム制御を取り入れ，論理プロトコルは FC を継承することでエンタープライズの SAN 環境からの移行性が高い．ここが同じく物理層に Ethernet を利用する iSCSI と異なる．FC では NPIV（N Port ID Virtualization）という WWN（World Wide Name）を仮想化する技術も出てきており，ゲスト OS の移行を行うサーバ仮想化環境，FCoE との組合せでの普及が期待される．

SCSI の標準化を進める ANSI（American National Standards Institute）T 10 では SCSI ブロックコマンド仕様の次期バージョン SBC-3 において，容量仮想化を適用したボリュームに対する操作管理コマンド（サーバ上で削除された領域をストレージに通知し，ストレージでその領域へのページ割当ての解除を可能とする操作など）を議論している．このような動きにより，サーバ仮想化でのリソース管理とストレージでのリソース管理が連携し，更に利用効率の高い運用が可能となる．SCSI だけでなく，サーバ管理ソフトが提供する API（Application Interface）を通じた連携も進展する動きがある [8]．

■参考文献

1) M. Palankar, A. Onibokun, A. Iamnitchi, M. Ripeanu,“Amazon S3 for Science Grids: a Viable Solution?”
2) “Hadoop,” http://hadoop.apache.org/
3) Shun-Tak Leung, “The Google File System,” SOSP 03.
4) D. Colarelli, D. Grunwald, “Massive Arryas of Idle Disks For Storage Archives,” ICS 02.
5) http://blogs.sun.com/blueprints/entry/running_sysbench_benchmark_on_mysql
6) “SUN Black box PJ,” http://jp.sun.com/products/sunmd/s20/features.html）
7) K. Li, “Aboiding the Disk Bottleneck in the Data Domain Deduplication File System,” FAST 08.
8) http://www.vmware.com/jp/products/vstorage-thin-provisioning/