高性能・高信頼ディスクアレイ装置
High-Performance and High-Reliability Disk Array Units

生藤　芳一　IKITO Yoshikazu
笹本　享一　SASAMOTO Kyoichi

昨今，PCサーバシステムやUNIX(注1)サーバシステムで，ディスクアレイ装置が適用される割合が増えてきている。特に，PCサーバは今後，性能の向上に伴い適用されるシステム規模が拡大していき，ディスクアレイ装置の適用率がますます増えていく傾向にある。ディスクアレイ装置は高信頼性と高速性を兼ね備えており，PCサーバシステムにおいては，データ保全とシステムの稼働率向上のために欠かせない存在になっている。当社は独自の高信頼化技術と高速化技術を用い，ディスクアレイ装置ArrayFort™シリーズを開発し，商品化している。

In recent years, there has been increasing demand for the incorporation of disk array units in PC server and UNIX server systems. Especially in the case of PC server systems, with their improvement in performance the system size is becoming larger and the proportion of systems incorporating a disk array unit is showing an increasing trend. Disk array units offer both high reliability and high performance, and are therefore an essential component of a PC server system in order to realize data integrity and high availability of the system.

Toshiba has developed and released the ArrayFort™ series of disk array units, featuring Toshiba's original high-reliability technology and high-performance implementation technology.

1 まえがき

ディスクアレイ装置とは，複数の磁気ディスクドライブ(HDD)を使用し，高速性と高信頼性を実現するRAID(Redundant Arrays of Inexpensive(Independent) Disks)技術を使った磁気ディスク装置である。

ディスクアレイ装置はサーバ内蔵のタイプと，外付けタイプに大別されるが，昨今，UNIXサーバシステムやPCサーバシステムでは，システム規模の拡大に伴ない，外付けタイプが増えてきている。当社は，UNIXサーバやPCサーバで使用できるオープンな外付けタイプのディスクアレイ装置ArrayFort™シリーズ(以降，ArrayFort™と略記)を開発，商品化し1997年から販売を開始している(図1)。

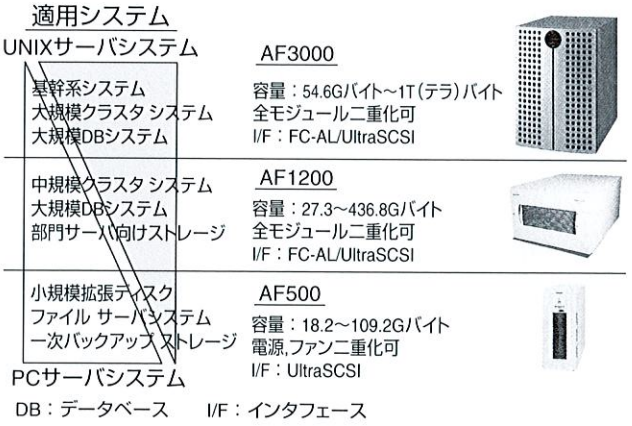

図1．ArrayFort™シリーズのラインアップ　小規模グループサーバから大規模基幹系システムまで，幅広いニーズにこたえることができる。
Lineup of ArrayFort™ series

2 RAID技術について

RAIDは，複数のディスクを組み合せて並列に動作させることにより読み書きの高速化を図り，データを冗長化(二重化や多重化)することにより耐障害性を上げる技術である。

RAIDは，RAB(The RAID Advisory Board)という標準化団体によりレベル0から6まで定義づけられており，一般に商用として有効なレベル0，1，3，5が実現されている。

レベル0はデータを複数のディスクに分けるストライピング方式，レベル1は同一データを2台のディスクに書き込むミラーリング方式，レベル3は1台のディスクをパリティ(誤り訂正符号)用のディスクとする方式，レベル5はレベル3のパリティを分散させる方式である。

ArrayFort™は，サーバシステムで要求が多いRAIDレベル0，1，5，及び1と0を組み合わせた1＋0をサポートしている。また，1台の装置の中で，これら複数のRAIDレベルのアレーを設定することが可能で，システムに応じてRAIDレベルを使い分けることができる。

(注1)　UNIXは，The Open Groupの米国及びその他の国における登録商標。

ArrayFort$_{TM}$は，ホットスペアディスク(注2)もサポート可能であり，1台のディスクが故障しても，データをホットスペアディスク上に復元させることにより，再び元の状態で安全に稼動させることができる(図2)。

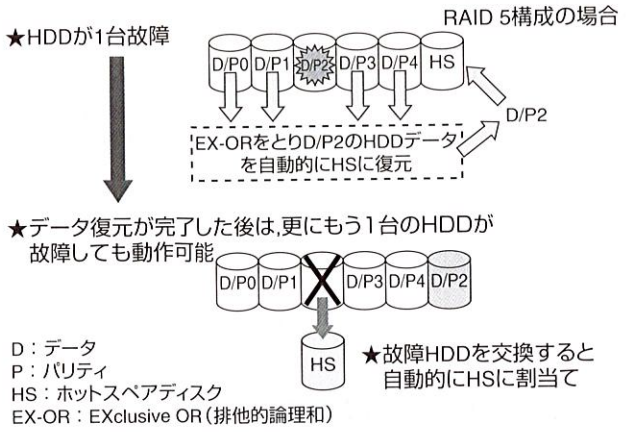

図2．ホットスペアディスクの動作　1台のディスクが故障しても，その内容をホットスペアディスク上に自動的に復元する。
Operation of hot-spare disk

3 高信頼化技術

RAID技術を使用することにより，HDDの障害対策をすることはできるが，それだけでは装置全体の信頼性を向上させることはできない。ディスクアレイ装置はディスクドライブのほか，コントローラ，ファン，電源から構成されるが，これらのモジュールについても信頼性を上げるために，冗長化という手段が採られている。AF500はファンと電源，AF1200とAF3000はコントローラを含むすべてのモジュールが二重化でき，高い信頼性を実現している。更に，保守面では，前面，背面からすべてのモジュールの交換ができ，しかも，サーバ側ソフトウェアの操作なしで，オンライン交換及び復旧が可能である。

また，ArrayFort$_{TM}$には次のような装置内部での高信頼化機能が搭載されている。

3.1 HDD耐障害性向上機能

一般的に，HDDの記録密度が上昇するに伴ないメディアエラーの発生率は増加していく傾向にある。“HDD耐障害性向上機能”は，ディスクアレイ装置運用時にオンラインで，また，並行してバックグラウンドでメディアエラーを検知し，エラー修復を行うことにより，装置の可用性(注3)を向上させる機能である(図3)。この機能の詳細について次に述べる。

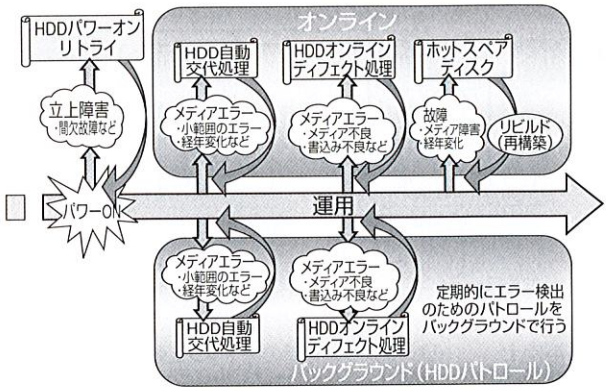

図3．HDD耐障害性向上機能　ディスクドライブのエラーを未然に防ぐことにより，システムの稼働率を向上させている。
Fault defending function for disk drive

(1) パワーオンリトライ　電源投入時に，正常に立ち上がらなかったHDDに対して，ドライブの電源を入れ直すことにより，再立ち上げを試みる機能

(2) オンラインディフェクト処理　オンラインでアクセス時に，あるドライブでメディアエラーが生じた場合，一般的には該当ドライブを切り離すが，ArrayFort$_{TM}$は同ディスクに代替え領域を設けるディフェクティブ処理をオンラインで施し，ドライブをすぐには切り離さず，継続して使えるようにする機能

(3) ディスクパトロール機能　バックグランドで装置内の全ドライブに対して定期的な診断を実施し，不良ブロックがあった場合は自動的に代替え領域を設ける機能

3.2 バッテリバックアップ機能

電源故障や停電に備え，ディスクキャッシュ上のデータを保持するバッテリバックアップ機能を搭載している。バッテリで保持されたデータは復電後にディスクに保存され，システム再起動時にもデータの整合性を保つことが可能となっている。また，バッテリモジュール自身もオンライン交換可能となっており，消耗時にも装置を停止させずに交換することができる。

3.3 ミラードキャッシュ機能(注4)

コントローラが二重化されている場合，一方のコントローラが故障すると他方のコントローラに切り換わるが，ディスクキャッシュ上のデータは常に相互にコピーされており，データが失われずに連続運転が可能である。ディスクキャッシュへのデータの書込みは，相手コントローラ上のディスクキャッシュにもデータが書き込まれてから完了するような機構になっており，データの整合性を保っている(図4)。

(注2)　あるディスクが故障した場合，故障したディスクのデータをパリティディスクから復元して，スペアディスク上にコピーすることにより，故障ディスクの代わりにすることができる。このスペアディスクをホットスペアディスクと呼ぶ。

(注3)　エラー時における稼働率の高さを示す指標。

(注4)　二重化された各コントローラには，キャッシュメモリがある。一方のコントローラが故障した場合を想定し，キャッシュメモリ上のデータを保持するため，データをミラーリング(鏡のようにコピーすること)している。これをミラードキャッシュと呼ぶ。

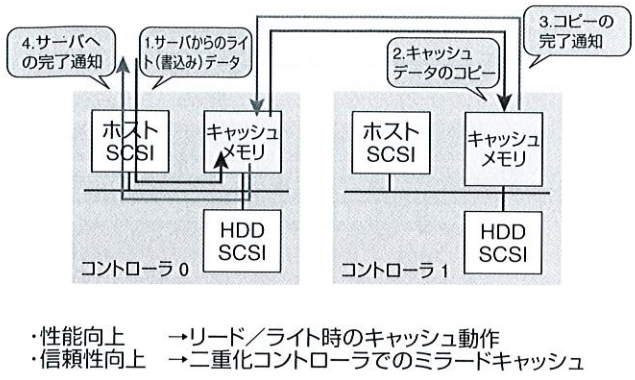

図4. ミラードキャッシュの機能　二重化されたもう一方のキャッシュにデータを書き込んでから，ライト動作を完了する。
Operation of mirrored cache

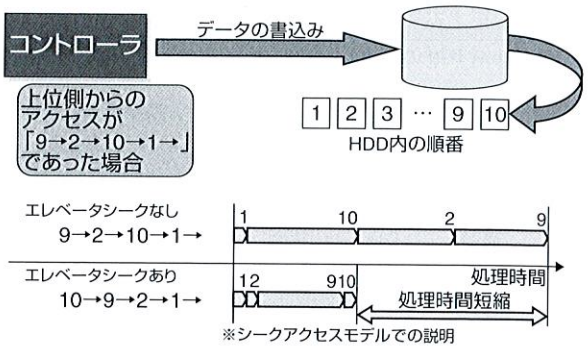

図5. エレベータシークによる高速アクセス　ランダムなアクセスをアドレス順に並び替えてアクセスすることにより，高速化を図っている。
Quick access by elevator seek function

4 高速化技術

ArrayFort$_{TM}$は，高い信頼性と同時に，性能についても独自の技術で業界トップレベルを実現している。

4.1 高速デュアルRAIDプロセッサ

AF3000ではRAIDプロセッサ（RAID処理装置）として，専用のゲートアレイを開発し採用している。RAID5時のパリティデータの生成やミラードキャッシュの制御などを二つのRAIDプロセッサで並列に処理することにより，性能を従来機の2～3倍向上させている。

4.2 大容量ディスクキャッシュ

ディスクキャッシュを搭載することにより，サーバからのリード／ライト時間を短縮することができる。AF500で64Mバイト/コントローラ，AF1200で最大128Mバイト/コントローラ，AF3000では最大2Gバイト/コントローラのディスクキャッシュが搭載でき，高いレスポンス性能を実現している。

また，AF3000ではリードキャッシュ(注5)とライトキャッシュ(注6)を同一のキャッシュ領域に割り当てることにより，ヒット率(注7)を向上させ，性能を上げている。

4.3 エレベータシーク

エレベータシークは，ランダムアクセスを高速化するための技術である。一般に，サーバでトランザクション処理(注8)を行うと，ディスクへのアクセスはランダムな領域へのアクセスとなることが多い。複数のランダムなアクセスを，データのあるディスクの順番に並び換えることにより，より短時間で処理することができる。これは，エレベータに乗った行き先の違う複数の人々を搭乗順ではなく，階数順に降ろす制御に似ている（図5）。

4.4 並列動作可能な複数の内部SCSIバス

複数のディスクへ並列にアクセスし，システム性能を向上させるために，装置の内部では，複数のSCSI（Small Computer System Interface）バスでHDDを接続している。AF500では2本，AF1200とAF3000では6本のSCSIバスを用い，同一アレイ内で並列にHDDをアクセスすることができる。

5 共有ディスクとしての機能

クラスタ システムとは，複数のサーバを使用して，システムの稼働率を向上させたり，並列処理によりシステム性能を向上させるものである。ArrayFort$_{TM}$は，このクラスタ システムを構築するための共有ディスクとして必要な機能を備えている。システムで扱うデータは共有ディスクに保存することにより，複数のサーバからアクセスすることができる。サーバ間で，アクセス時に適宜排他(注9)を行うことにより，スタンバイシステムではデータの引継ぎを行い，並列処理システムではデータの共有を行うことができる。

ArrayFort$_{TM}$は次のような共有ディスクとしての機能を持つ。

5.1 独立動作可能なマルチポート インタフェース

AF1200とAF3000には，それぞれのコントローラに複数のホスト接続用のインタフェース ポートを持つ。SCSIタイプではコントローラごとに四つのSCSIポートを持ち，Fibre Channel(注10)タイプではコントローラごとに二つのFC-AL（Fibre Channel–Arbitrated Loop）(注11)ポートを持つ。

これにより，SCSIタイプでは同時に四台のサーバと接続が可能であり，Fibre ChannelタイプではHUB(注12)を使うこと

(注5)　サーバからデータをリード（読込み）の場合に有効になるキャッシュ。
(注6)　サーバからデータをライト（書込み）の場合に有効になるキャッシュ。
(注7)　サーバからアクセスした際に，キャッシュ上にデータがある（ヒットする）確率のこと。確率が高いほど性能が良くなる。
(注8)　一つの意味のある処理単位をトランザクションと呼び，ここではサーバからの一連のデータアクセス処理のことを言う。
(注9)　あるサーバがアクセスしている場合は，他のサーバからはアクセスしないようにする機能。
(注10)　最大転送速度100Mバイト/sの高速インタフェース規格。
(注11)　Fibre Channelの接続形態の一種で，機器間をループ状態に接続し，複数の機器からアクセスが同時に重なった場合は，規格に沿って調停（Arbitration）が行われる。
(注12)　FCのネットワークに接続するための機器。

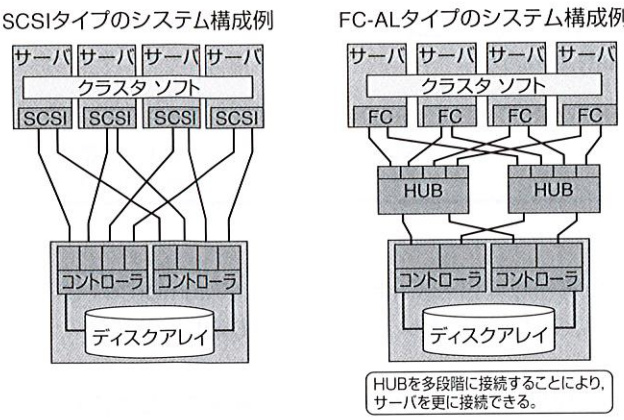

図6．独立動作可能なマルチポート インタフェース　マルチポート インタフェースにより，複数のサーバから同時にアクセスを受け付けることができる。
Multiport interface supporting simultaneous access

により，最大126台のサーバと接続が可能である（図6）。

また，これらのポートは独立動作が可能となっており，複数のサーバから同時にアクセスがある並列処理システムにおいて高いシステムスループットが実現されている。

5.2 コントローラのリザーブ／リリース機能

スタンバイシステムでは，共有ディスク上のデータを保護するために，運用系のサーバからのアクセスは許可し，待機系のサーバからのアクセスは許可しない排他制御を一般的にはサーバ上のクラスタソフト(注13)が実行している。

(注13)　クラスタ システムを構築するときに使用するソフトウェア。

ArrayFort™では当社クラスタ ソフトDNCWARE™と連携し，ハードレベルでリザーブ／リリース機能による排他を実施しており，データの保全性を高めている。

6 あとがき

今後，コンピュータシステムにおいて，ますますディスクアレイ装置の役割が重要になってくる。簡単にかつ柔軟に高信頼システムが構築でき，また安心して手間を掛けずに運用できる商品が望まれている。自動性能チューニング機能や他メディアへの自動バックアップ機能など運用性を意識した機能，障害時にも短時間でデータを復旧する機能について今後取り組んでいく。当社は，ArrayFort™シリーズで更なる高信頼性と高速性を追求していくとともに，ユーザーにとって最適なシステム構築環境と各種サポートサービスを提供していく計画である。

生藤　芳一　IKITO Yoshikazu
デジタルメディア機器社 コンピュータ・ネットワーク事業部 コンピュータ・ネットワーク商品企画担当主務。ディスクアレイ装置の商品企画業務に従事。情報処理学会会員。
Computer & Network Div.

笹本　享一　SASAMOTO Kyoichi
デジタルメディア機器社 府中デジタルメディア工場 コンピュータハードウェア部主務。ディスクアレイ装置の開発に従事。情報処理学会会員。
Fuchu Operations－Digital Media Equipment