

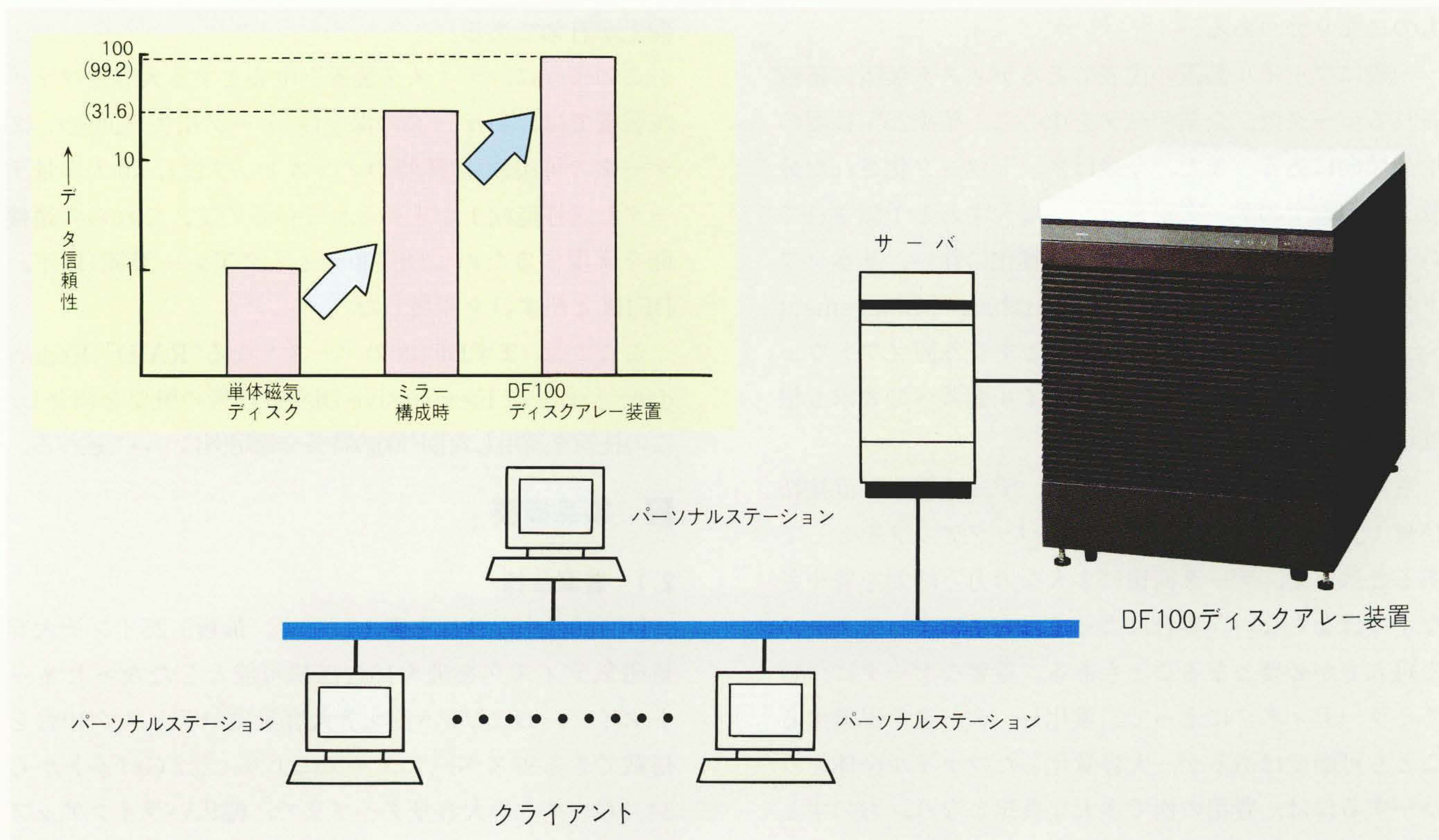
高速データ転送性能とデータ保全性能を兼ね備えた 小型ディスクアレー装置

Small Disk Array Unit with High Data Transfer Speed and High Data Security

星野政行* Masayuki Hoshino

村岡健司* Kenji Muraoka

永井英男* Hideo Nagai



システム構成例 小型ディスクアレー装置をデータベースサーバシステムに適用する際のシステム構成例を示す。左上のグラフは、単体磁気ディスクおよび単体磁気ディスクでのミラー構成時とのデータ信頼性の比較を示す。DF100ディスクアレーは、同一容量であればミラー構成と比較して約3倍のデータ信頼性を持つ。

近年、半導体技術の発達、およびCPU性能の目覚ましい向上により、ワークステーション、中・小型コンピュータ、さらにはパーソナルコンピュータなどの高性能化が急速に進みつつある。これに伴い、オープンシステム、クライアントサーバシステムといった分散情報システム、および高速・大規模ネットワークシステムの発達が目覚ましい。このような状況下で、ワークステーション用ファイルサーバ、およびPC/LANサーバ用の外部記憶装置として、大容量、高速転送、かつ高信頼性を特長とするDF100デ

ィスクアレー装置(以下、DF100と略す。)を開発した。

DF100は、小型磁気ディスク装置を複数台使用し、アレーコントローラで並列処理することにより、大容量かつ高速データ転送を実現した。また、冗長データを自動生成付加することにより、1台の磁気ディスクが故障しても、ユーザーデータを損失することなくデータを回復させることができる。さらに、従来の大型汎(はん)用装置で培われたデータ保証技術を採用し、ユーザーデータの信頼性をより高めている。

* 日立製作所 ストレージシステム事業部

1 はじめに

情報システムのダウンサイジング化、およびネットワーク化に伴い、処理の分散化、データの共有化が図られるなど、情報処理システムの利用形態は多様化している。このような背景のもとで、パーソナルコンピュータ(以下、パソコンと略す。)、ワークステーションでのファイルサーバなどのファイル装置の役割も、ますます重要なものになりつつある。

一般にファイル装置の代表であるディスク装置に蓄積されるデータは、企業データを中心に、年率25%程度の増加傾向にある。また、今後はネットワーク化された分散システムでのデータが飛躍的に増大すると予測されている。ソフトウェアの高度化・複雑化に伴い、基本ソフト(OS)やRDBMS(Relational Database Management System)などミドルウェアを代表とする各種ソフトウェアも増加しつつあり、大容量ファイル装置への要求も根強い。

さらに、ファイル装置の故障は、データ損失に直接結び付く。通常はテープ装置などでバックアップをとってあるとはいえ、データ回復に多大な労力と時間を費やさなくてはならない。場合によっては人手によってデータの再入力が必要となることもある。重要なデータについてミラーディスクによって二重化し、データを保護することも可能ではあるが、大容量化したファイル全体をカバーするには、費用の面で多大な負担となり、おのずと限界がある。特に、オンラインシステムでは、ディスク

装置の故障が直ちにシステムダウンとなってしまう。これも小容量データでは、ミラーディスクによって保護できるが、大容量データ全体を二重化することは困難である。

画像データ、科学技術計算などの大容量データ転送時では、その転送データ量が膨大となる場合が多く、データ処理時間のうちファイル装置にかかわる部分が多い。このような場合、ディスク装置のデータ転送速度の向上が有効である。

このように、ディスク装置を中心とする大容量ファイル装置では、(1)データの保全性(データ損失の回避)、(2)データの可用性(故障時のノンストップ性)、(3)大容量データ転送性能向上が求められているので、これらの諸機能を実現するためにDF100ディスクアレー装置(以下、DF100と略す。)を開発した。

ここでは、まずDF100のベースとなる“RAID”(Redundant Array of Inexpensive Disks)技術の概要を紹介し、この技術を適用したDF100の特長や適用例について述べる。

2 製品概要

2.1 装置仕様

DF100の主な仕様を表1に示す。最新5.25インチ大容量磁気ディスクを最大16台搭載可能としたキャビネットタイプ、および3.5インチ大容量磁気ディスク10台を搭載できる省スペースタイプにより、2.2 Gバイトから34.4 Gバイトの大容量タイプまで、幅広いラインアップを図った。また、高性能RISC(Reduced Instruction Set

表1 DF100の主な仕様 省スペースタイプ、キャビネットタイプによって2.2 Gバイトから34.4 Gバイトのラインアップを図った。

項目		型式		DF100-2	DF100-4	DF100-5	DF100-11	DF100-12	DF100-17	DF100-34
記 憶 容 量				2.2 Gバイト	4.5 Gバイト	5.6 Gバイト	11.2 Gバイト	11.4 Gバイト	17.2 Gバイト	34.4 Gバイト
イン タ フ ェ ース				SCSI-2〔10 Mバイト/s(Fast), 20 Mバイト/s(Wide)〕						
ド ラ イ ブ 数				5	10	5	10	5	8 スペアディスク 1台を含む。	16 スペアディスク 2台を含む。
内蔵 デ ィ ス ク	デ ィ ス ク 装 置			3.5インチ薄型		3.5インチ		5.25インチ		
	記 憶 容 量			573 Mバイト		1.43 Gバイト		2.87 Gバイト		
	転 送 速 度			2.7～4.5 Mバイト/s				4.8 Mバイト/s		
	平均シーク時間			12.4 ms		11.8 ms		12.8 ms		
外 形 寸 法 (幅×奥行き×高さ)				250×600×600(mm)					570×800×700(mm)	
タ イ プ				省スペースタイプ					キャビネットタイプ	
質 量				最大70 kg					最大200 kg	
電 源 条 件				AC100-120 VまたはAC200-240 V+ 6 %/－10%					AC100 VまたはAC115 Vまたは AC220-240 V+ 6 %/－10%	
消 費 電 力				最大700 VA					最大1,500 VA	

注：略語説明 SCSI-2(Small Computer System Interface-2)

Computer)プロセッサ採用による高速処理能力に加え、業界標準であるSCSI-2 (Small Computer System Interface-2) インタフェースをサポートし、10 Mバイト/s (Fast SCSI) および、20 Mバイト/s (Wide SCSI) の高速データ転送を実現した。

さらに、RAID 5 構成により、データ消失防止など耐障害性を飛躍的に向上させるとともに、高トランザクション性能をも実現した。

2.2 主な特長

(1) 高信頼性

RAID 5 構成をサポートし、磁気ディスク 1 台に故障が発生してもパリティデータがデータを保護するので、大切な情報資産を失うことなく連続して稼動させることができる。従来の単体磁気ディスクに比較して、データ損失の可能性を数百分の一に低減した。

(a) 制御部：従来大型汎(はん)用機だけに使用されていたデータ保証コード、およびホストインタフェース部入口から磁気ディスクまでのスルーパリティにより、データの信頼性を向上させた。

(b) 磁気ディスク：ヘッド・ディスクからの自社一貫生産を行った。高信頼性部品、専用LSIの採用、および厳格な品質管理によって高信頼性を誇っている。

(c) 電源部：キャビネットタイプでは、大型汎用機で実績のあるCVT(定電圧トランス)を採用した。

(d) 保守診断プログラム：保守診断および障害対策時に威力を発揮する保守ポートを標準装備し、保守ポートには、業界標準インタフェースであるRS-232Cを採用し、パソコンからも保守診断を可能とした。

(2) SCSI-2 標準インタフェース

業界標準のSCSI-2 インタフェースをサポートする。従来のSCSIとの完全上位互換で、さらに高度なコマンド処理機能を合わせ持ち、ワークステーション、パソコンをはじめ、各種システムに適応する汎用性を備えている(シングルエンド、ディファレンシャル、ワイド、ファーストの各種SCSIをサポート)。

(3) 高性能

高速磁気ディスクの採用およびこれら資源を最大限に生かすソフトウェア・ハードウェア構成により、高パフォーマンスを実現した。また、高性能RISCプロセッサおよび高性能SCSI-2の採用により、20 Mバイト/s (Wide SCSI) または10 Mバイト/s (Fast SCSI) の連続データ転送能力を持ち、特に大量データ処理に威力を発揮する。

(4) 小型・大容量

最新の5.25・3.5インチ大容量磁気ディスクを、キャビネットサイズでは16台、省スペースタイプでは10台実装可能である。2.2 Gバイトから34.4 Gバイトの記憶容量を実現した。

(5) 豊富な機能

ホストとの間で安定したデータ転送を行うディファレンシャルSCSI、保守管理、障害対策時に威力を発揮するRS-232C保守ポートなど、大容量磁気ディスク装置にふさわしい機能を装備している。

2.3 ハードウェア構成

DF100(省スペースタイプ)の内部構造を図1に示す。

DF100は磁気ディスクをアレー状に配置、接続したドライブ部、それら磁気ディスクを制御するアレー制御部、ファンユニット、およびそれらに電源を供給する電源部で構成する。

(1) プラグイン脱着方式HDUの採用

障害ディスクの交換作業をより容易にするため、磁気ディスクドライブをモジュール化構造(HDU: Hard Disk Unit)とした。また、ケーブル類の使用を極力避けて、HDUの脱着はプラグイン方式を採用し、HDUの交換は前面だけで可能とした(キャビネットタイプは前・後面を使用)。

さらに、保守作業者が誤って障害HDUと異なるHDUの脱着操作を行わないように、パネル上に障害HDUの位置表示を行うとともに、保守ポートに接続されたパソコン画面上にも、位置表示可能とした。特に、省スペース

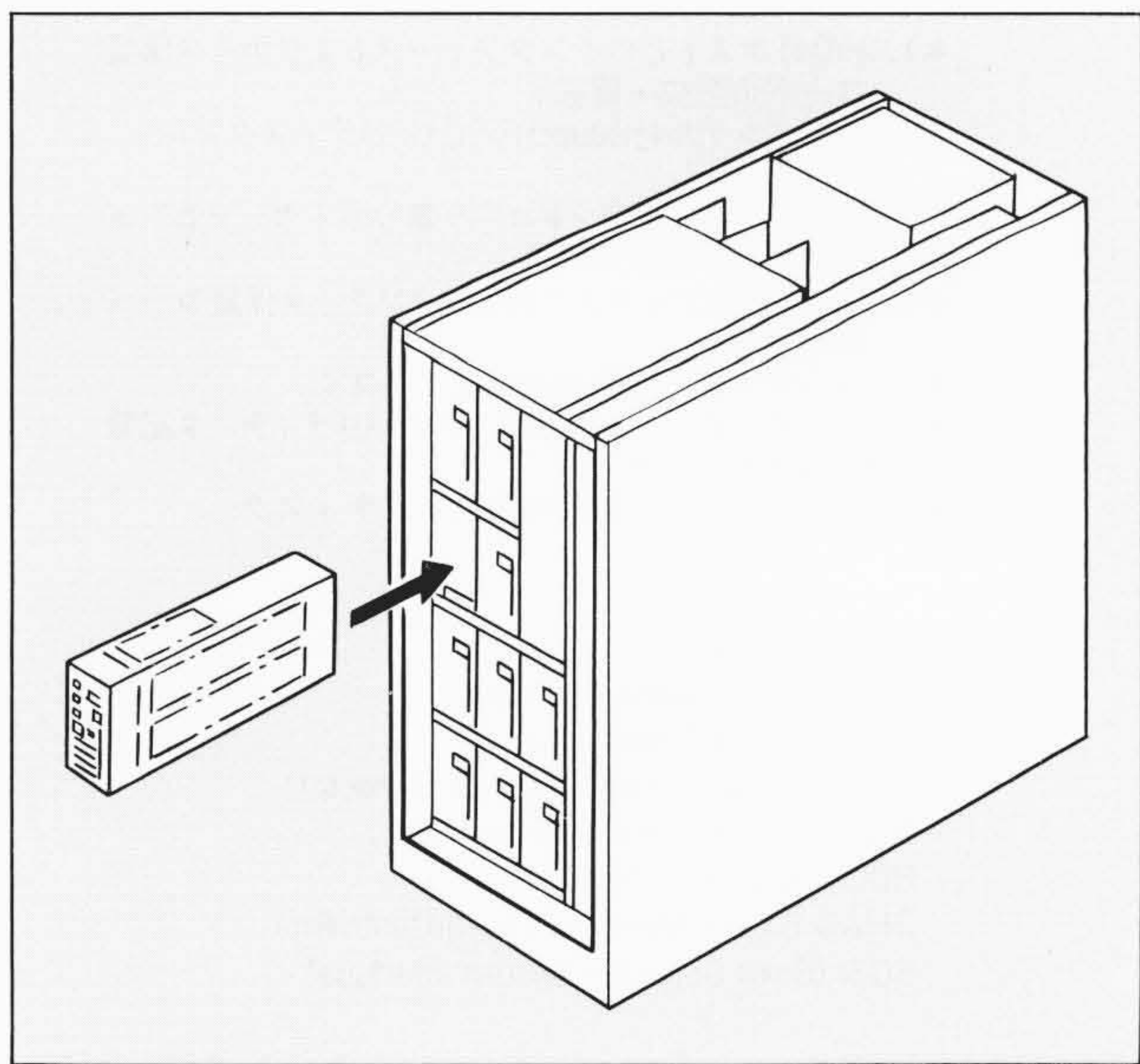


図1 DF100(省スペースタイプ)の内部構造 動作中でも障害発生ディスクを容易に交換できる実装を実現した。

タイプでは、HDUに障害を示すLED(Light Emitting Diode)を設け、障害HDUを一目で判別可能とし、作業性の向上を図った。

(2) 連続高速データ転送および高速多重処理性能を実現

DF100の制御ブロック図を図2に示す。DF100制御部は、MPU部、HIC(Host Interface Control)部、PIC(Peripheral Interface Control)部、およびDMA/BUF(Direct Memory Access/Buffer)部の四つの部位で構成する。

MPU部は、装置全体の制御を行う。メインプロセッサとして、32ビットRISCプロセッサ(米国インテル社製i80960CA)を持ち、RAID特有のデータの集合・分散演算処理の高速化を実現しているとともに、HIC部およびPIC部のサブプロセッサとコミュニケーションをとりながら、多重処理のスケジューリングを行っている。

HIC部は、上位ホストインタフェースであるSCSI-2を制御し、ホストとのデータ転送制御を行う。特にデー

タ集合・分散制御用LSIで、上位ホストとの転送制御をはじめ、パリティデータ生成回路への同時転送制御をハードウェア化することにより、上位ホストと20 Mバイト/sでの連続データ転送を可能とした。

PIC部は、下位ドライブインタフェースであるSCSIを制御し、ドライブとのデータ転送制御を行う。特に転送制御用LSIは、磁気ディスク上に分散書き込まれたパリティデータの読み落とし制御、データの部分書換動作時に必要となるダミーデータ転送制御、さらには、データ修復時のパリティデータ生成回路への同時転送などの機能を持ち、ハードウェア制御で高速化を実現している。

HIC部およびPIC部には、それぞれサブプロセッサとして、16ビットCPU(米国インテル社製80C186XL)を持ち、メインプロセッサで展開されたサブコマンドを個々に実行制御可能としている。これにより、RAID 5構成での磁気ディスクの多重処理性能の向上を図っている。

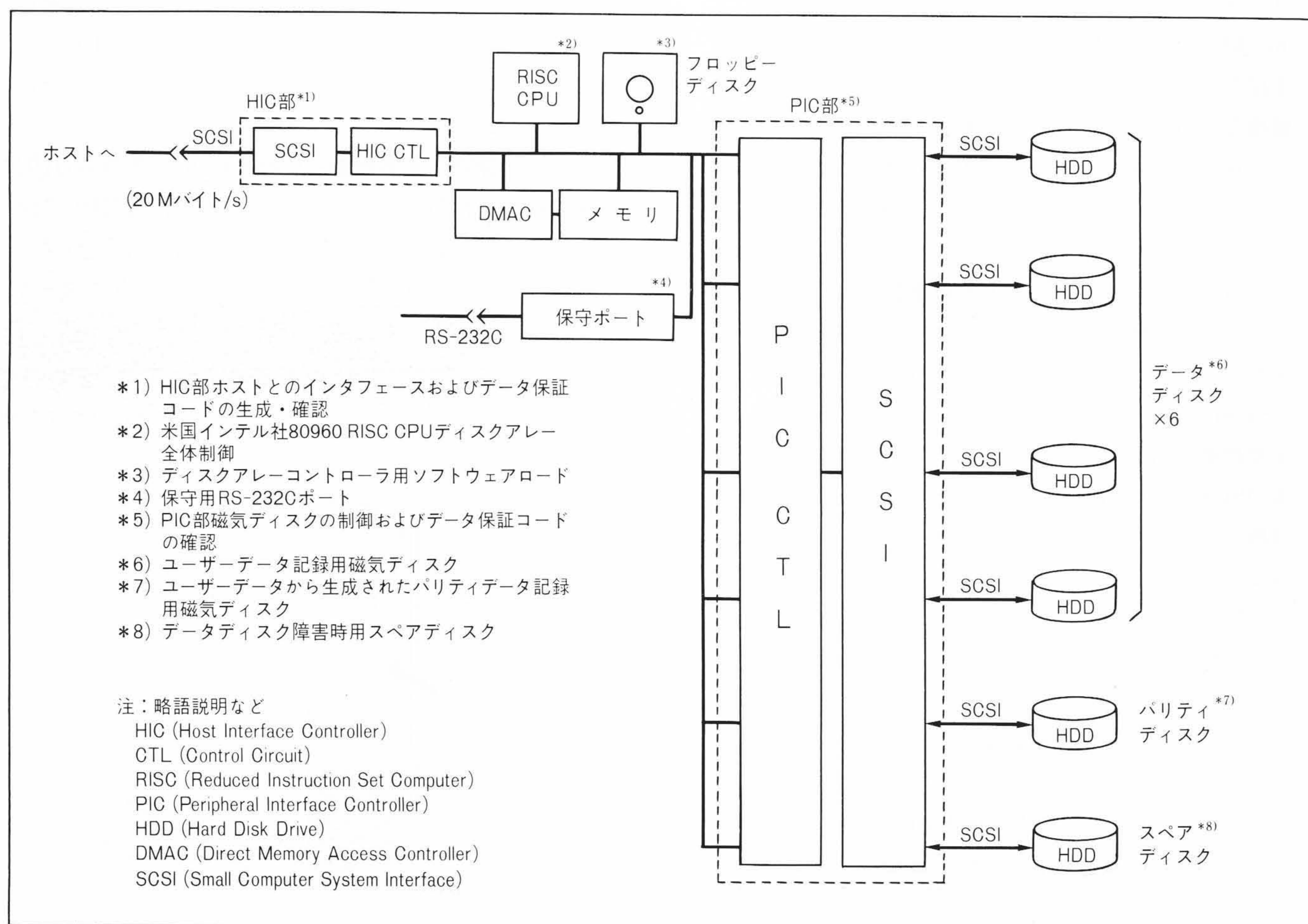


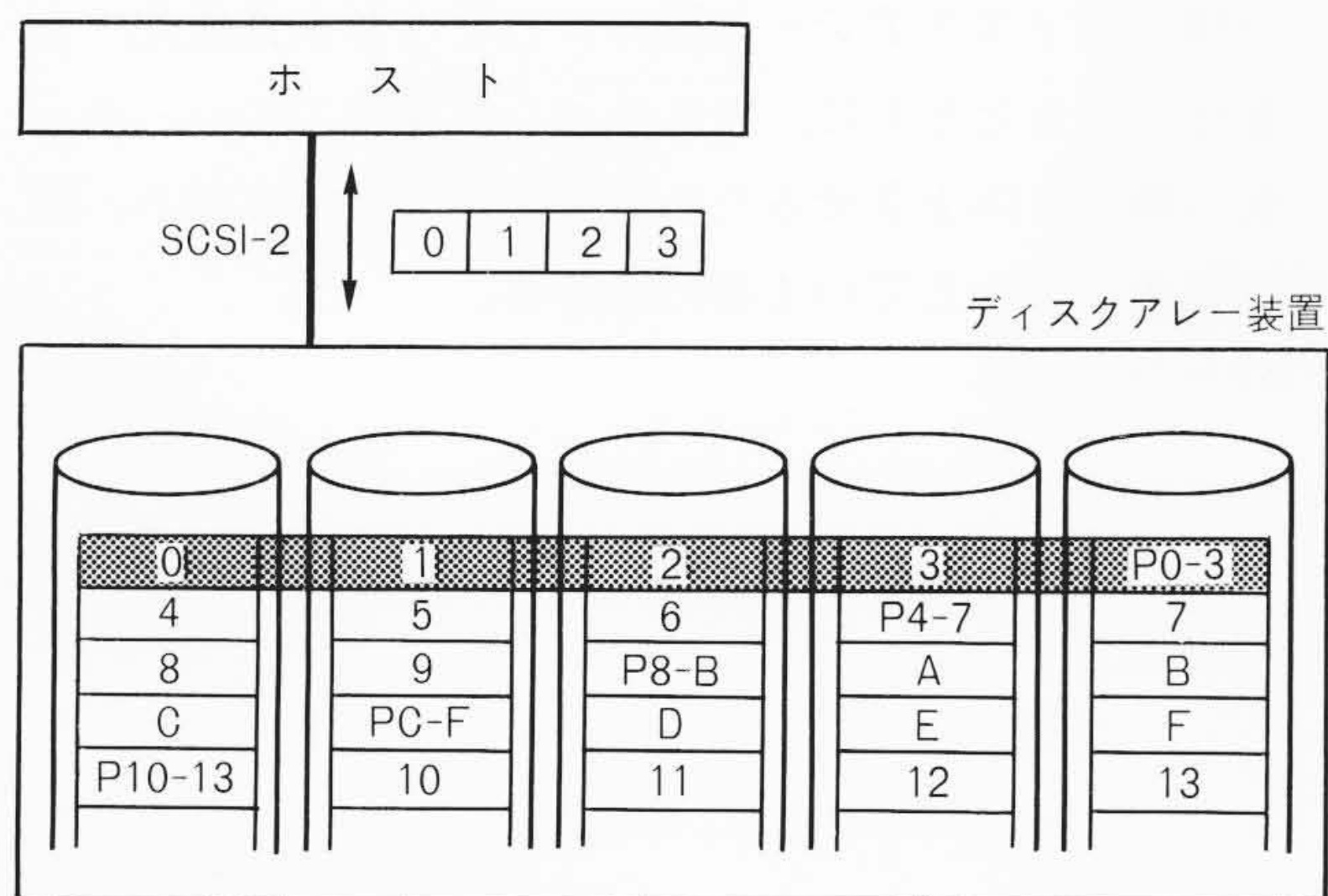
図2 DF100のブロック図 RISCプロセッサ、サブプロセッサの使用とデータ転送制御回路のハードウェア化により、アレー装置特有のオーバーヘッドを最小限に抑える構成とした。

DMA/BUF部は、データの集合・分散のためのデータバッファの管理、およびHIC部、PIC部から送られるデータの集中転送制御を行う。DMA部の機能としては、内部データバス(50 Mバイト/s)のクロック制御、MPU部、HIC部、PIC部およびパリティデータ生成回路部からのバス使用権獲得要求の調停を行う。また、バッファのフルエンpty制御、および分散データ読出し時の自動頭出しなどをハードウェア制御化している。さらに、パリティデータ生成部での生成データの自動転送制御、および同一パリティグループ以外のデータ誤転送抑止制御など、ハードウェア化によってプロセッサへの負荷を最小限にしている。

(3) 制御部から磁気ディスクまで一貫したデータ保証

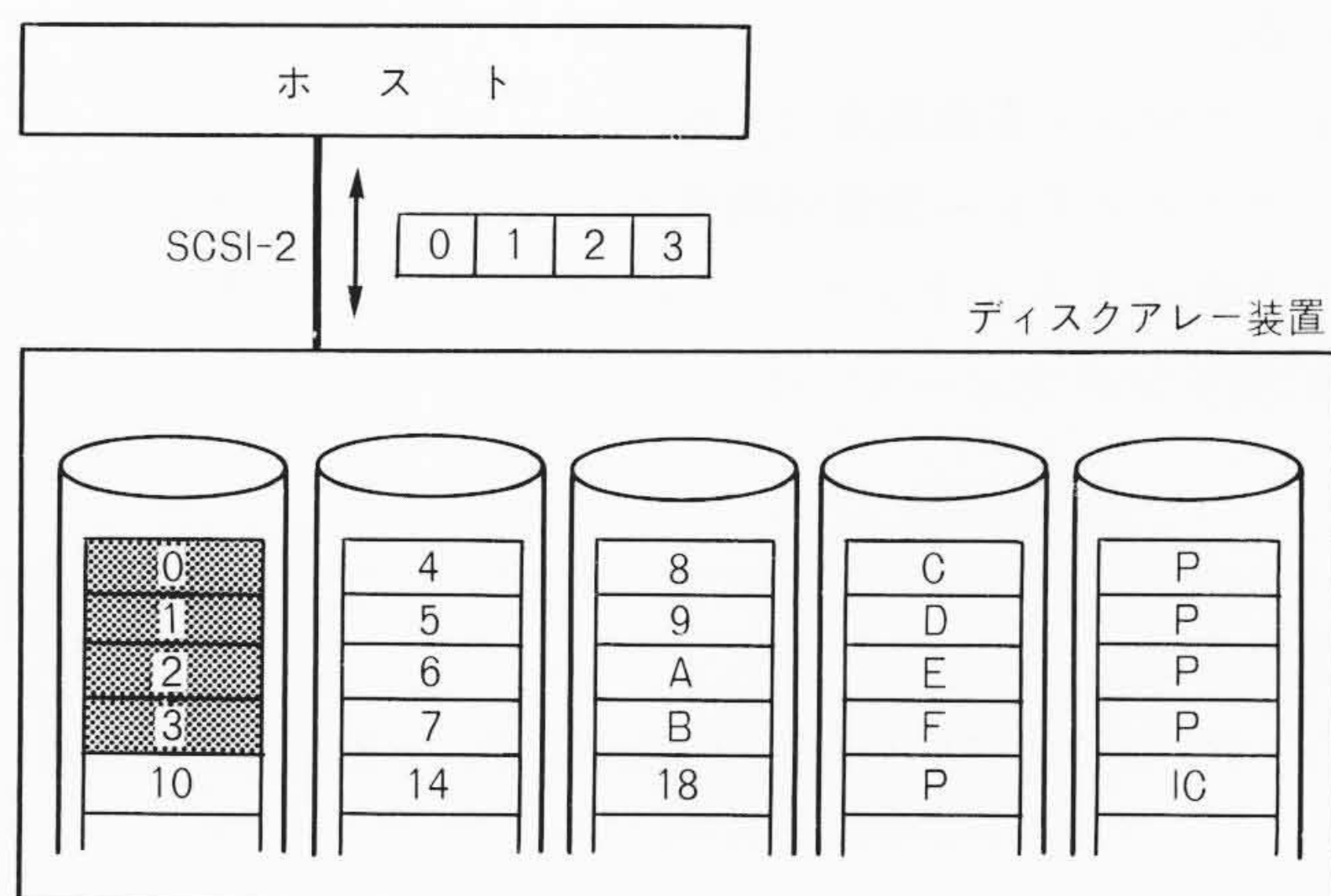
ディスクアレーコントローラは、一般的にデータの集合および分散処理を行うため、従来の磁気ディスクコントローラ以上のデータ信頼性保証が必要である。特に一連のデータ群を複数台の磁気ディスクに分散し、書込み、さらにはパリティデータをも磁気ディスク上に割り付けなければならない。このようなデータ処理過程で、いかにデータ自身および書込み位置を保証できるかが重要なポイントとなる。

DF100では、上位ホストインタフェース部入口から磁気ディスクに至るまでのスルーパリティに加え、データ自身を保証するLRC(Logical Block Address)、書込み位置を保証するLBA(Longitudinal Redundant Code)などの保証コードを付加してデータの信頼性を高めている。



■ : ストライピング単位, P: パリティ, 数値: ブロックアドレスストライプ長を小さくとり、ホストデータを各単体ディスクに分散させることによってホストデータ転送性能を向上する。

(a)



■ : ストライピング単位, P: パリティ, 数値: ブロックアドレスストライプ長を大きくとり、1I/Oを1台の単体ディスクに収めることによって多重処理性能を向上する。

(b)

図3 DF100でのデータ分散方式 大容量データ転送向けデータ配置を(a)に、トランザクション処理向けデータ配置を(b)に示す。

3 マルチベンダ環境システム

情報システムのダウンサイジング化は、マルチベンダ環境でのシステム構築とも換言でき、ディスクアレー装置でも各種ホストシステムとの親和性・接続性が要求されている。このホストシステムの親和性・接続性を実現するために、DF100で実現した先進的な技術要素について以下に述べる。

3.1 標準ディスクデバイスドライバによるホストシステムへの接続の実現

ディスクアレー装置をホストシステムに接続する際に、最初に問題となる点がデバイスドライバの扱いである。DF100では、ホストインタフェースとしてSCSI-2を採用し、コマンド仕様は単体ディスクと互換性を持たせることにより、ディスクアレー装置専用のデバイスドライバを不要とし、各ホストシステムでの標準のディスクデバイスドライバによる接続を可能にした。これにより、ホストシステムへの接続時にデバイスドライバの組み込みなどの操作を不要とし、種々のホストシステムへの容易な接続を可能としている。

3.2 多様なシステム運用形態への対応

単体ディスクと比較したディスクアレー装置での性能上の優位性は、

- (1) ディスクドライブ並列データ転送によるホストデータ転送性能の向上
- (2) コマンド多重処理によるトランザクション性能の向上

である。(1)の高速データ転送を実現するためには、ストライプ長(ホストデータをディスクアレー内の単体ディスクに振り分ける単位)を小さくとり、ホストデータを各単体ディスクに分散させる必要がある〔図3(a)参照〕。また、(2)の高トランザクション性能を実現するためには、ストライプ長を大きくとりホストからの1回のI/Oを1台の単体ディスク内に収め、各単体ディスクがそれぞれ別のホストコマンドを実行できるようにする必要がある〔同図(b)参照〕。このようなシステムのさまざまな運用形態に対応できるように、DF100では、ストライプ長を最小512バイトから最大32kバイトの範囲で設定可能としている。

3.3 論理ボリューム分割機能の提供

マルチベンダ環境で使用されるホストシステムは多種多様であり、各ホストシステムによってディスクのサポート仕様は異なっている。DF100では以下の二つの問題の解決のために論理ボリューム指定機能を提供した。論理ボリューム指定機能は、1台のDF100をSCSIインタフェースで規定されている複数のロジカルユニットに分割し、各ホストシステムから個々のロジカルユニットを1個の独立したディスク装置として取り扱い可能とすることによって実現している。

(1) 使用可能ディスク容量の拡張

ディスクアレー装置の特長の一つに大きな記憶容量の提供があるが、ホストシステムによってはディスク装置1台当たりの最大サポート記憶容量が制限されており、DF100のサポートしている最大記憶容量を扱えないケースでは、複数論理ボリュームへの分割によって解決している。

(2) コマンド多重処理の実現

ディスクアレー装置の特長の一つにコマンド多重処理の実現によるトランザクション性能の向上があるが、SCSIインタフェースでは、タグドコマンドキューイング

を使用しなければ1台のディスク装置に複数のコマンドを同時に発行できない。しかし、タグドコマンドキューイングをサポートしているホストシステムはまだ少ない。そのため、複数論理ボリュームへの分割の支援によってタグドコマンドキューイング未支援のホストでもコマンド多重処理ができるようにした。

3.4 媒体エラーフリーの実現

媒体上のきずによるリード・ライトエラーが発生した場合、通常、交代ブロックによる復旧が必要となる。DF100では媒体エラー検出時に自動的に交代ブロックを設定し、さらにエラーによって読出し不能となったブロックのデータもパリティデータによって自動的に回復することにより、媒体エラーフリーを実現してホストシステムでのエラー回復の負担を低減した。

3.5 ノンストップ障害ディスク交換の実現

DF100内蔵の単体ディスクで障害発生時に、障害ディスクの交換および交換したディスクへのデータの再構築は、ホストからのコマンド処理のバックグラウンドで行うことにより、ホストシステムを停止させることなく障害ディスクの交換を可能とし、可用性を向上させた。

4 おわりに

マルチベンダ環境システム向けのファイル装置に対する大容量化・高信頼性などの幅広いニーズに対応するため、DF100を開発した。このディスクアレー装置は高性能RISCプロセッサを採用し、大型汎用機向けディスク装置で培った技術によってデータの信頼性・可用性を高めている。

今後、ディスクアレー装置のいっそうの高性能化・高信頼化を図るとともに、マルチベンダ環境でのいっそうの使い勝手を向上させるための技術開発に取り組み、顧客の要求にこたえていく考えである。

参考文献

- 1) ディスク・アレー装置：性能向上で、分散システムの要に、日経エレクトロニクス、No.579(1993.4.26)