

パネル討論:

ディスクアレイの現状と展望

司会: 喜連川 優 (東大)

パネリスト: 金子 悟 (富士通)、中村 俊一郎 (三菱)、
山本 彰 (日立)、辻澤 隆彦 (日本電気)

従来プロセッサの性能は著しい向上を遂げてきたのに対し二次記憶装置とりわけディスクは大容量化、小型化に関しては進歩があるもののそのアクセスタイム、トランスファレートなど性能面では殆ど改善されていない。この第2次フォン・ノイマンボトルネックを回避すべく最近ディスクアレイの開発が活気を帯びてきている。現時点でその多くはRAID3 であり、アプリケーションが特化しているものの、今後大きなマーケットが期待できる。本パネルではディスクアレイに関するアーキテクチャ、制御方式、アプリケーション、コストなど種々の側面から現状と今後について模索してみる。

Panel Discussion:

Disk Array : Present and Its Perspective

Chaired by: Masaru Kitsuregawa (University of Tokyo)

Panelists: Satoru KANEKO (Fujitsu), Syunichirou NAKAMURA (Mitsubishi),
Akira YAMAMOTO (Hitachi), Takahiko TSUJISAWA (NEC)

The MIPS rate of CPU has been so far improved dramatically, on the other hand, the performance of secondary storage devices, such as seek time and transfer rate has not changed so much. In order to overcome this 2nd von-Neumann Bottleneck, recently the research and development of disk array become hot issues. At present, most of the current systems belong to RAID-3 device. But the market seems to grow rapidly. In this panel, we pain to discuss several topics of the disk array such as architecture, control strategy, applicationcost etc. and derive some perspective.

□ディスクアレイの動機付け

1. I/O ボトルネックを何とかしたい

I/O ボトルネックの問題が顕著になり始めて以来久しいが、充分なアプローチがとられてきたとはいえない。PTD やSilicon Diskなど2次記憶のアクセスタイム、転送レートを上げる手法や拡張記憶に於けるハイパー空間など、ソフト的にもI/O への配慮が一段と増している。特にメインレーム大型機に於いてはわずかではあるが、チャンネルのスピードを向上させたりエスコンなどI/O 効率の向上への配慮がうかがえる。そもそもこれらのディマンドはユーザの急速に巨大化するデータベース、バッチアプリに負うところが大きいといえよう(e.g. DB2)。今や、CPUの改善は先がみえており、I/O がターゲットとなりつつある。

2. 小型ディスクの低廉化が火種

ディスクの容量の大型化はDRAMチップの大容量化と同様に大きな進歩をとげてきた。しかし一方でパソコンやラップトップなど超小型化に対する潜在要求が強く、小型ディスクの低価格化が進み、これらを沢山並べた方がコスト的にも性能的にも大型ディスクに比べ、優れる様になってきた。

3. 新しいアプリケーションも高速I/O を要求

画像処理、ミニスーパーコンなどミッドレンジの超高速アプリユーザ層が急増しつつある。

従来CRAYなどhighend はPTD を使用していたが、中速マシン(例、FPS, Maspar)などでの潜在ニーズが大きい。Array Techはタンデムに買収されたもののR-5 の有効性は未だ明確とはいえない。

□テクノロジー

ディスクをアレイ化すること自体は何も新しいテクノロジーではない。動画を扱うアプリケーションでは数十台のディスクの並列動作は過去にもプロダクトとに例がある(PEL)。

現在、R-1, R-2, R-3, R-4, R-5 (R-6, R-7)が模索されている。

R-1, ミラーディスク

R-2, データボルト

R-3, 各社開発中

R-4, 意味なし

R-5, OLTP

R-1: とにかくミラーは安心だし、シャドーイングの効果もある。

R-3: 最も手固いが、従来のSoftwareストライプと比べてどうか? スケーラビリティはあるか? など疑問も残るが、みんな作っているから...

R-5: コミットコントロール、リカバリーなど大変
(R-2, R-4は殆ど意味なし)。

□マーケット

CIPRICO	R-3	NCR	R-5
CORE	R-3	AMPERIF	R-5
DELL	R-3	AUSPEC	R-5
MICROPOLIS	R-5	SF2	R-5
STORAGE COMP	R-7	STORAGETEK	R-5
ARRAY TECH	R-0, 1, 3, 5	IBM	R-3
MAX	R-3, 5		

パネル討論：ディスクアレイの現状と展望

金子 悟 (富士通株式会社)

ディスクアレイとは、複数のディスクで構成するサブシステムであり、UCBerkeleyのPattersonらによって、RAIDRedundantArrays of Inexpensive Disksと呼ぶ5種類の形態に分類されている。ここでは、その現状について説明し、次に将来への展望を述べる。

1. 現状

ここ数年、ディスクアレイの開発は多くの企業で行われてきた。RAID-1:Mirrorは従来からソフトウェアで実現されていたが、RAID-3:Parallel Transferのものが主に製品化されている。RAID-3は、高速転送、高信頼性にその特長があり、RAIDの効果が最も分かりやすい。表-1に富士通の製品例を示す。想定するアプリケーションは、科学技術計算、CAD、画像処理等の高速データ転送である。

2. 将来展望

大容量ディスクサブシステムの構成は、将来的にはディスクアレイが標準になることが予想される。その理由として、以下が挙げられる。

(1) テクノロジー

今後の記録密度の向上に伴い、適正なデバイス記憶容量のために、ディスクドライブは小径化していく。また、小径化は高速回転および、高速ポジショニングのためにも必須である。

(2) 性能

一台あたりの容量の大きなディスクで構成したサブシステムよりも小容量のディスクを多数並べた方が、性能的に有利である。またデータ転送速度が必要な場合でも、複数台のディスクを並べて並列転送させた方が、特殊な高速転送ディスクよりもコスト的に有利である。

(3) 信頼性

大容量のディスクサブシステムでは、信頼性が重要であり、何らかの冗長性が必須となる。二重化構成はその点に関しては十分であるが、大容量になればなるほど、コストが気になってくる。RAID-3,4,5のパリティディスクは、n台に1台の冗長性で実用上問題にならない信頼性(MTTL)を確保することができる。

RAIDの各レベルは次のようになると予想している。

RAID-1: Mirror

大容量、大規模構成になるとドライブコストの負担が大きいが、制御法がシンプルなので余計なコントローラコストがからず、小規模構成ではむしろパリティディスクの冗長性より有利である。従って、将来においてもある範囲で使われ続けられる。

RAID-3 : Parallel Transfer

RAID-3では、並列転送化してデータ転送を高速化したことにより、接続インタフェ

ースをSCSI-2や光ケーブルインタフェースにしなければならない。また並列ディスク分をまとめた論理セクタ長が標準より長くなるために、OSの対応が必須になる。性能を出すためには一度にまとまった量のデータ転送を行なわなければならないため、アプリケーションの対応も必要となる。このようにRAID-3はやや普通のディスク環境からはずれるために、メインのディスクシステムと言うより、その特性である、高速転送を活かすアプリケーションのためのディスクシステムである。しかし効果が分かりやすいため、先ずRAID-3から製品化されるのが自然の流れである。

RAID-5 : Spread Data/Parity

RAID-5はRAID-3とは逆に、ディスクの

環境は個々のディスクと変わらないため、メインのディスクシステムになりうる。但し、現在のところライトで不利になる性能上の欠点を克服したアルゴリズムは未だ登場していない。またそのようなコントローラはきっと複雑な構造となる可能性が高く、コスト面での課題も残るだろう。しかし楽観的に考えれば現在はRAID-5の一つの形態が提案されているだけであり、まだまだ異なる形態が登場してもおかしくない。もともと大規模構成を前提に考えるならば、キャッシュなど他の技術と組み合わせて使うのが自然であり、そのような中で解決策を見いだすべきであろう。

	大型用 ディスクアレイ	小型用 ディスクアレイ
データディスク容量 基本 最大	15GB 120GB	2GB 4GB
インタフェース	高速光チャネル	Fast SCSI-2
データ転送速度	36MB/s	10MB/s
平均ポジショニング時間	12ms	6.8ms
平均回転待ち時間	6.9ms	12ms
セクター長	4096Byte	2048Byte*
RAID構成	8+P+S	4+P(+S)

*1024Byteから設定可能

表－1 ディスクアレイ製品の仕様

パネル討論 ディスクアレイの現状と展望

— RAIDレベル4、5の考察 —

中村 俊一郎 三菱電機情報電子研究所

5. 2.5インチ、3.5インチといった小型ディスクが容量／性能面で急速に伸びてきており、これらを多数並べたディスクアレイが最近注目を集めている。又これには市販のLSIの技術的進歩が大きく係わっているという点も見逃せない。

ここではRAIDレベル4と5にしぼってその実現上の技術的問題点に対する考察を行なう。

1. 狭義のRAID

RAID (Redundant Array of Inexpensive Disks) は、単に従来方式のディスクをアレイ状に並べただけのRAIDレベル0と呼ばれるものから、二重故障までを復元可能なRAIDレベル6と呼ばれるものまで広い意味で使われることがある。RAIDレベル1は所謂ミラーディスクのことであり、ディスクの信頼性を向上させる目的で、既に多くの商用マシン上で使われている。RAIDレベル2と3についても、スーパーコンピュータ用とか、画像データのような大量のデータの高速転送用として以前から商用化されている。これに対しここ数年注目を集め初めてきた所謂狭義のRAIDとは、RAIDレベル4、5、6のことである。この狭義のRAIDについては、ミラーディスクのように多くの冗長ディスク(100%)を持たなくてもディスクのエラー訂正が出来るという魅力を持っている反面、ディスクへのWriteに時間がかかるという欠点も持っており、その評価はまだ固まった段階にはない。ここ数年米国では10数社がこれらのRAID製品を発表しているが、国内メーカーからはまだ本格的なRAID製品は出ていない。

2. RAIDレベル4、5

RAIDレベル2、3は複数のディスクを同期回転させて、データを並行に読み書きして高速転送を実現したものであるが、レベル3では1バイトを単位として各ディスクからデータ転送を行なう。RAIDレベル4は原理的にはRAIDレベル3のバイト単位を、ブロック(1～数セクター)単位にしただけの違いしかない。しかしながらブロックになると、単1ブロックのみの読み書きが可能になるためかなり変わった様相を呈してくる。ディスク

の同期回転とか複数同時アクセスは必要無くなり(即ち個々のディスクへの単独アクセスが可能となり)、レベル3では当たり前であったデータのストライピング(横並び)も必ずしも必要でなくなり、又パリティの生成方法等も変わってくる。RAIDレベル5はレベル4からパリティを全ディスクに分散させるという改良を加えたものであり、本質的にはレベル4と同じである。

3. WRITE性能

ディスクのエラー訂正のためにミラーディスクでは冗長ディスク量100%を必要とするが、RAIDレベル4、5では9～25%程度でよいため非常に魅力的である(前者は多重故障、後者は単一故障という違いは有るが)。その代わり図1に示されるようにRAIDレベル4、5ではディスクへのWRITE時には常に、①旧データREAD、②旧パリティREAD、③新データWRITE、④新パリティWRITEと従来型に比べ4倍のディスクアクセスが必要であり大きな性能低下を招く。これは換言すればディスクが故障したときのための保険料として、ミラーディスクは冗長ディスク量で払い、RAIDはディスクWRITE時の処理量で払うということができる。

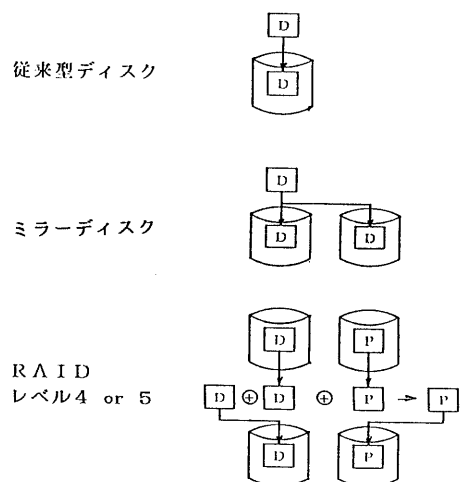


図1 ディスクのWRITE処理

4. L S I の進歩

上記のようにRAIDレベル4、5ではディスクへのWRITE 処理が遅いという大きな欠点を持っているが、この改善策としては以下のようなものが考えられる。

- (1)データディスクとパリティディスクを別のDKC(Disk Controller)に繋げ処理を並行して行なう。
- (2)データとパリティをキャッシュに入れ、上記①旧データREAD、②旧パリティREADをディスクキャッシュから行なうようにする。

さてここで注目すべきことは、Pacstor 社のIntegralのような初期の製品を除き、現在市販されているRAID製品はほとんど例外なく上記(1)を実現していることである。即ち5～12個程度のDKCを配置しディスクへのREAD/WRITEを並行処理して性能向上を図っている。しかもこの5～12個のDKCが1枚のカードにコンパクトに収められているため手頃な価格で実現出来る。従来それ自体1つの装置であったDKCが、複数個まとめて1枚のカードに収まってしまうというようなLSI技術の進歩があつて初めて、多数のDKCが必要とされるRAIDが商用化可能となってきたと言える。なおここでDKCとはディスクのデータ転送を独立して実行出来る単位のことを言っており、具体的にはSCSI I/O Processorを指す。(往年のDKCが備えていた機能の多くは、LSIの進歩により今やディスク装置自体の中に入っている。)

さてこれらの装置を眺めると「逆の面」が見えてくることに気付く。即ち1枚のカードに収まった5～12個のDKCが並行して処理を行なう訳であるから、RAIDのWRITE処理の遅さということ差し引いても、ディスクシステム全体としてかなりの高性能を発揮するだろうということである。このことが現時点においてRAIDの大きな魅力となっていることは否めない。

5. ストライピングと深さ

前章で述べた「ディスクシステムの性能」を考える時、それは以下の2点に集約される。

- (1)ディスク処理のスループット
(トランザクション処理の高速化)
- (2)ディスク処理の応答時間
(大量連続データREAD/WRITEの高速化)

さて、2章で述べたようにRAIDレベル4、5ではデータのストライピングは必ずしも必要ないが

、ストライピング無しの場合には、性能面では従来型のDKCが5～12個付いている場合に帰着でき、後はRAIDのWRITEによる性能低下を考慮すれば良い。

ストライピングをする場合には、RAIDレベル3型と従来型をミックスした使い方となるため、新たな技術的問題が生起する。ストライピングの効果は以下のようなものである。

- (1)データベースマシン等で使われる水平分割手法と同様、1つの論理的なファイルが自動的に複数のディスクにばらまかれるため、特定のディスクにアクセスが集中してネックになることを解消できる。
- (2)ディスク上の連続データを大量にREAD/WRITEする場合にはすべてのDKCが並行してデータ転送を行なうため高速転送が可能となる。一方次のようなデメリットも存在する。
 - ①個々のディスクのスケジューリングはRAID側が行わざるを得なくなり、それに伴いO/S側にも変更が必要となる。
 - ②上記の(2)使い方をした場合、連続データが大量でなかった場合に却って性能(スループット)の低下を招く。例えば連続データが数KByteで3個程度のディスクに渡るようなケースでは、1個のディスクに収まった場合に比べて明らかにスループットが落ちる。このことはストライピングの深さの問題に帰着され、これはユーザが選択可能なパラメータであるが重要な戦略的意味を持つことになる。

6. まとめ

以上主としてRAIDレベル4と5の技術的問題点につき考察を行なった。今後益々システムが分散化し、その中心となるサーバ上のストレージシステムの重要性は増加し、この種の議論が活発になることを期待したい。

[参考文献]

- (1)D.A.Patterson et al. 'A Case for Redundant Arrays of Inexpensive Disks(RAID)' Report No. UCB/CSD 87/391, 1987
- (2)PRODUCT DESCRIPTION' RAID+SERIES MODEL RX' Array Technology Corporation, 1990

パネル討論：ディスクアレイの現状と展望

山本 彰 (株)日立製作所システム開発研究所)

1. ディスクアレイの現状について

まず、ディスクアレイの定義であるが、ここでは、ユーザから見た1つの論理ディスク上のデータを、複数の物理ディスクに配置したものと考え、単なる2重書きディスクは除くものとする。ディスクアレイのタイプには、大きくいうと、高速データ転送専用タイプと汎用ディスクの置き換えも可能な高トラフィックタイプに分類できる。もちろん、高トラフィックタイプでも、ユーザから見た1つの論理ディスク上のデータを、複数の物理ディスクに配置しているのであるから、ディスク間並列転送により、従来のディスクに比較し、大幅な転送速度向上が期待できる。

ディスクアレイの開発は、最近急速に進んでおり、製品化、あるいは、その発表を行っているメーカは非常に多くなってきている。ただし、製品化されているほとんどのディスクアレイは、高速データ転送専用タイプである。また、高トラフィックタイプでは、STK社のアイスバーグが注目されている。

2. ディスクアレイの展望

結論から述べると、ニーズから考え非常に有望であると考えている。理由を以下にまとめる。

(1) 高速データ転送に対するニーズ・・動画など本質的に高速性を要するマルチメディアなどの新アプリケーション、オンライン時間拡大に伴うバッチ時間の圧迫等から、大量のデータを高速転送するニーズが高い。しかし、ディスク装置単体の転送速度の向上は、このニーズに追従していないため、ディスクアレイのようなアプローチを取らざるを得ない。

(2) 信頼性の向上に対するニーズ・・現在の計算機システムでは、ディスク装置がファイルの恒久的な格納媒体となっているため、信頼性に対するニーズは極めて高い。従来、ディスク装置の信頼性の向上のためには、2重書きというアプローチが取られた。しかし、2重書きの場合、信頼性は飛躍的に高まるが、2倍のディスク台数が必要となる。これに対し、ディスクアレイの場合、m台に対し1台分の冗長データを設定することが可能となり、ディスク台数を段階的に増やしていくことにより、信頼性も段階的に向上させることができる。さらに、冗長データを2重化することにより、2重書きの場合と比較し、少ない増設ディスクで、2重書き以上の信頼性を得ることができる。

以上が、ディスクアレイが今後有望と考えられる点である。反面、汎用ディスクに比較すると以下に示すような短所がある。

(1) ライト処理に対するペナルティ・・信頼性の向上のため、冗長データを設けた場合、データの内容の書き換えに伴い、冗長データの更新値の作成と更新値の書き込み処理が発生する。

(2) CPU側から見た入出力処理の実行並列度の低下・・同時に、複数のディスク装置を占有するため、CPU側から見た入出力処理の実行可能な並列度が低下する。

しかし、今後、マルチメディア処理等のディスクアレイの特徴が活かせるアプリケーションが増大することが予想され、普及は進むと考えられる。

パネル討論：ディスクアレイの現状と展望

辻澤 隆彦 (日本電気(株)機能エレクトロニクス研究所)

1. はじめに

小型ディスク装置を使ったディスクアレイ装置は、ディスク装置単体でのドラスティックなパフォーマンス向上があまり期待できなくなった現状から、飛躍的な性能向上の一つの解決策として注目され、メインフレームだけではなく、サードベンダからも製品化が進められるようになってきた。それには以下の様な理由が考えられる。

- ①大型ディスク装置に比べ、アクセスストローク/消費電力等の点で有利な小型ディスク装置が、大型ディスク装置と技術的に遜色がなくなってきたこと。
- ②小型ディスク装置のアレイ化で大容量化が可能であり、ビット単価を低下できること。
- ③データ転送レートを向上させることができること。
- ④データの信頼性を向上させることが可能であること。

しかしながら、市場の立ち上がりから見ると④の理由による製品化が支配的である。データストライピングによる高速転送レートディスクアレイ装置の実用化には、ストライピングによる制御の複雑さやオーバーヘッドの増大等の効果的解決が残されている。

ディスクアレイ装置の特徴は既に示したように高速化と高信頼性にある。従って、実用化を考えるに当たっては、この二つの特徴をどの様にバランスさせるかが重要であり、このためにもアプリケーションをにらんだ装置の実用化が必要である。本文では、この観点からディスクアレイについて概観してみたい。

2. ディスクアレイアーキテクチャ (RAID) と適応アプリケーション

ディスクアレイには5つのレベルがあることは周知の通りである。また、ディスクアレイは既に述べたように高速性と高信頼性を提供するが、どのようなRAIDアーキテクチャを採用するかによってアプリケーション領域はほぼ決まってしまう。RAIDアーキテクチャではRAID-1、3、5が代表的であり以下これらを例に適応アプリケ

ーションと特徴を記す。

(1) RAID-1 (ディスクミラーリング、ディスクデュプレッシング)

ディスクアレイの中でもっとも実用化が進んでいる。方式的には送られてきたデータをマスタディスクアレイに書き込むと同時に予備ディスクアレイにも書き込むといったフォルトトレラントを主目的にしたディスクアレイアーキテクチャである。ディスク装置4台程度で実現されることが多く、LANサーバ (特に、PCサーバ) における実用化が進んでいる。

LANサーバでの負荷集中回避と信頼性向上を目的にI/Oコントロールを多重化するディスクデュプレッシングも同時に進められている。

(2) RAID-3 (パリティ専用ディスク)

RAID-1に比べRAID-3は、データ転送レートの向上を目的としたアーキテクチャである。この方式では送られてきたデータはバイト単位 (あるいはビット単位) に分割され、アレイ化されたディスクに並列に書き込まれる。データが書き込まれる際に分割されたデータからパリティを計算し、パリティ専用ディスクにこれを記録しておくことが特徴である。しかし、ディスク装置の物理的特性からデータ転送の高速性を保持するためには以下の関係を満足することが必要である。

$$BZ > SZ * N \quad (1)$$

BZ: データブロックサイズ (ホスト)

SZ: ディスク装置のセクタサイズ

N: アレイ化ディスクの数

アプリケーションとしては、転送ブロックサイズの大きい科学技術計算/画像処理用イメージサーバが考えられる。

(3) RAID-5 (パリティディスク)

RAID-5はオンライントランザクション処理のような比較的処理データの少ない場合への対応を考慮したアーキテクチャであり、データ格納方式としては、ホストから送られてきたデータを複数のディスクに対してブロック単位で順次保存していく方式を取る。全てのディスクドライブが同時に動作する必要がないために、

処理の多重化による高速化が図れる。しかし、読み出し動作に比べ書き込み動作でのオーバーヘッドが大きい。データベースの様な読み出しを中心としたアプリケーションでの利用価値が高い。

3. インテリジェントディスクアレイコントローラ

以上見てきたように、代表的RAIDアーキテクチャとアプリケーションはほぼ対応している。しかしながら、RAIDを実用化するに当たっては、RAIDレベルが上がると共にディスクアレイコントローラに要求される計算能力も大きくなることから、これまではRAID-3レベルのディスクアレイ装置の実用化が進められるにとどまっている。また、市場の魅力としてもPC、WS市場への展開が重要視され、今後ともRAID-3レベル以下の製品化が進められるものと考えられる。こうした動きの中で、活性交換/ディスクミラーリング/ディスクデュプレッシング機能を備えたもの、あるいは、アプリケーションに応じてRAID-1からRAID-5までを選択できるようなインテリジェントコントローラの開発が活発化しており、ディスクアレイ装置によるデータの高信頼性化は進められるものと考えている。

4. ディスクアレイとフルテキストサーチ

さて、著者らはこれまでに流通サービス業や金融向け接客サービスといった観点から、ディスクアレイ装置と文字列検索LSI (ISSP) を応用した接客端末装置の開発を進めてきた。ディスクアレイ装置の一つのアプリケーションとして捉えることができると考えここに紹介したい。

この装置はPC-H98をベースに、32ビットバス (NESA) に4台のディスクアレイ装置と文字列検索ボードを接続した構成になっている。

著者らは無人接客端末の基本機能として検索をとらえ、高速フルテキストサーチとイメージブラウジング機能をディスクアレイ装置により実現することを目標に端末開発を行った。ディスク装置はSCSIインターフェースを持つものを使った。このディスクアレイ装置ではディスク装置への書き込み/読みだしはSCSIインターフェース上のバッファとディスクコントローラ上のFIFOを介して行われるため、ディスクコントローラ側

がバッファリングを行うことなくディスクの非同期運転に対応している (図1)。(転送ブロック長が通常転送の5~6倍であれば、非同期転送による特性悪化を相対的に低減できる。3))

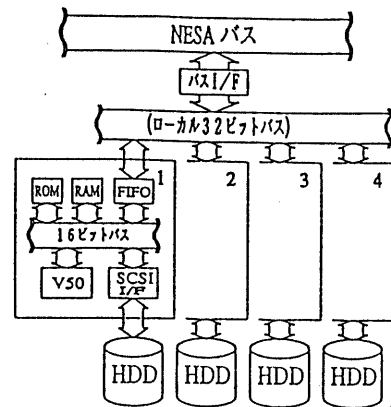


図1 ディスクアレイ構成

本ディスクアレイ装置による性能評価結果以下の通りである。

フルテキスト検索 : 約3.5 MB/s
イメージブラウジング : 約10枚/秒
(640 * 400 16色)

5. おわりに

ディスクアレイ装置をアプリケーションの観点から概観した。ディスクアレイ装置はPC-LANサーバ、ファイルサーバといったアプリケーション領域から市場が立ち上がりつつあり、RAID-3を中心に実用化が今後急速に進むものと思われる。また、これと同時にデータチャネルの高速化を狙ったHIPPI等のインターフェースの採用も進むであろう。さらには、ディスクアレイコントローラの高機能化はディスクアレイ装置構築を容易にしていけるものと思われる。

参考文献

- 1) Spencer Ng: "Some Issues of Disk Arrays", CO MPCON'89 Spring, 1989
- 2) 高密度記録媒体関連プロジェクト'91 調査報告書, (株)野村総研
- 3) 杉本、菊地、辻澤; マルチディスク装置(MD-1)の性能評価, 信学会春期全国大会予稿集, 1991
- 4) Steven J. Vaughan-Nichols; Disk Insurance, BYTE, AUG. 1991