Vol. 37 No. 11　　情　報　処　理　　Nov. 1996

解　説

計算機システムを支える最新技術（装置編）

2. ファイル装置の信頼性向上技術†

菊地芳秀††　辻澤隆彦††
右田守彦†††　赤木正信†††

1. まえがき

1988年に発表されたPattersonらの発表論文[1]を契機に，ファイル装置の信頼性向上技術はディスク単体の信頼性向上を目指す方向から，安価なディスクを用いながらもファイル装置全体の信頼性向上を目指す方向へと変化してきている[9]．

Pattersonらは論文の中でRAID（Redundant Arrays of Inexpensive Disks）という概念を提唱し，Level 1-Level 5に分類した．その考え方の基本は冗長性を埋め込むことにあり，代表的なものがRAID-1（Mirroring），RAID-3（Bit-Interleaved Parity）およびRAID-5（Block-Interleaved Distributed Parity）である[2]．

本稿では，まずRAIDを中心としたディスクアレイ技術について概説し，つづいてメインフレーム向けディスクアレイを例にとり信頼性向上技術の実例について紹介する．また，ディスクアレイのアプリケーションとして最近増えているリアルタイム系アプリケーションからみたディスクアレイへの要求についても簡単に触れる．

2. ディスクアレイ技術

ここではRAIDを中心とした信頼性向上技術について説明するとともに，性能問題となるライトペナルティについて述べる．

2.1 RAID技術[1)~5)]

図-1から図-6にRAIDのレベルを簡単に図示する．図において網かけ部分は冗長部分（パリティブロックなど）を示す．

† Technologies for Improving Data Reliability of Storage Systems by Yoshihide KIKUCHI, Takahiko TSUJISAWA (Functional Devices Labs., NEC), Morihiko MIGITA and Masanobu AKAGI (Storage Products Division, NEC).

†† NEC機能エレクトロニクス研究所
††† NECファイル装置事業部

（1） RAID-0（図-1）

RAID-0は説明上の分類で，信頼性の向上を目的としたものではない．複数のディスクを並列動作させることで入出力速度の向上を狙ったものである．

しかしながら，最近の小型ディスクの読み出し速度の向上は目覚ましく，2~3台接続するとSCSIインタフェースが飽和してしまうケースもみられるようになり，RAID-0の意義は薄れつつある．

（2） RAID-1（図-2）

RAID-1はミラーリングを行う方式であり，ディスクを2倍必要とする．2つのディスクに同じ内容を書き込み，読み出し時は早く読み出せる方から読み出すことにより高速化をはかる方式が一般的である．一方，両方のディスクから読み出した内容を比較することでデータの信頼性を高める方式をとるものもある．

（3） RAID-3（図-3）

RAID-3~6では，データを一定の長さに分割してディスクに順番に格納する．これをストライピングと呼ぶ．RAID-3では，ストライピングの単位を小さくし（たとえば1ビットや1バイト単位），リード/ライト要求に対して，同時にすべてのディスクへアクセスすることが特徴となっている．パリティディスクへの書き込みも同時に行われることから，リードもライトも同じスピードで処理できる．しかもディスク障害が起きた場合，復旧動作をさせなければリード/ライトの速度は通常時と変わらないというメリットがある．

RAID-3は，一般にはリード/ライトの単位が大きくなる点やアクセスの並列処理ができない点から，小規模データのアクセスに向かないとされている．ただし，複数のRAID-3ユニットを並列動作させることで同時アクセス数を増やし，ト

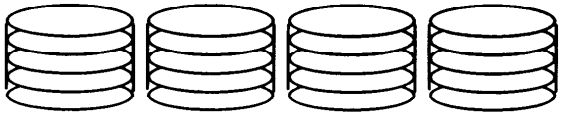

図-1 RAID-0

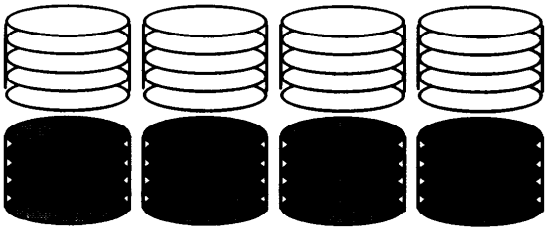

図-2 RAID-1 (Mirroring)

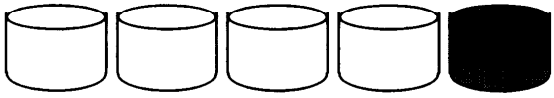

図-3 RAID-3 (Byte-Interleaved Parity)

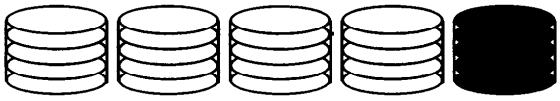

図-4 RAID-4 (Block-Interleaved Parity)

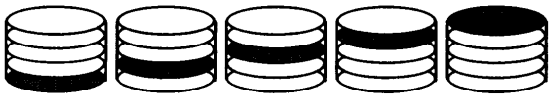

図-5 RAID-5
(Block-Interleaved Distributed Parity)

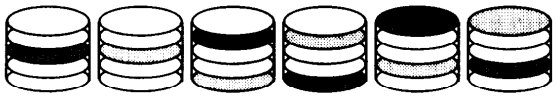

図-6 RAID-6 (P+Q Redundancy)

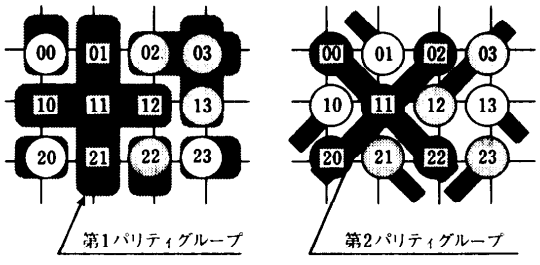

図-7 DR-Net によるディスクアレイ

ランザクション処理などの小規模データのリード/ライトに対応させるものもある.

(4) RAID-4 (図-4)

データ分割の単位を大きくした方法 (ブロック分割) が RAID-4 や RAID-5 である. ブロック分割では, 各ブロックの排他論理和 (XOR) をとった結果がパリティブロックに書き込まれる. このパリティブロックを同じドライブに固定したのが RAID-4 である.

後述するようにライトペナルティが重いため, RAID-4 は一般的には採用されない.

(5) RAID-5 (図-5)

RAID-4 がパリティブロックを1つのディスクに固定したのに対し, RAID-5 ではパリティブロックを全ディスクに分散しているのが特徴である. ライト時に起きるライトペナルティが各ディスクに分散されるため, RAID-4 よりもライトの並列処理を効率よく行うことができるという特徴をもつ.

(6) RAID-6[2),4)] (図-6)

基本的には, RAID-5 を拡張した方式である. 大きなディスクアレイシステムになるとパリティグループ内においても複数台のディスク障害が想定される. この方式は, リードソロモン符号化に基づいたパリティブロックを2つ用意し, パリティグループ内で最大2個のディスクの障害に対応できる. ライトペナルティは RAID-5 に比べ大きくなる.

2.2 メッシュ状のディスクアレイ

ディスクアレイを高速な内部バスで接続したものに, TickerTAIP/DataMesh[6)] がある. 小規模なディスクアレイと異なるのは, 複数のコントローラボードをもち, その間を高速な内部バスで接続している点である. コントローラやホストとのインタフェースの一部に障害が生じてもディスクのアクセスが可能になっている. 3章で説明するが, 現在のメインフレーム用ディスクアレイでは, ほとんどがこれに似た構造を採用して信頼性を上げている.

ディスクをネットワーク上のノードに接続し, ネットワーク上のサブネットでパリティグループを形成するものとして DR ネットワーク (Data Reconstruction Networks)[7)] がある. ディスクの複数台故障へ対応するために, 第1パリティグ

996

ループと第2パリティグループを構成し，この2種類のパリティグループを重ね合わせるという方式をとっている．

2.3 ライトペナルティ

RAID-3～6のディスクアレイのライトでは，目的のブロックの書き込みのほかにパリティブロックをも書き換える必要がある．一般にパリティブロックは，パリティグループに含まれている複数のデータブロックのXORをとって生成される．

RAID-3の場合，パリティグループに含まれるすべてのブロックを同時に書き換えるため，古いパリティブロックの情報は必要なく，ライトペナルティは起こらない．

しかしながらRAID-4～6では，1つのデータブロックとパリティブロックを書き換えるには，データブロックのリード/ライト，パリティブロックのリード/ライトと合計4回（RAID-6ではパリティブロックが1つ多いことから合計6回）のディスクアクセスが発生することになる．このことを一般にライトペナルティと呼んでいる．

RAID-4ではパリティブロックを1つのディスクに固定化しているため，ライトの集中時にパフォーマンスが低下する．この理由からRAID-4はあまり用いられない．

一方，RAID-5についてはホストからのライト動作とディスクに書き込む動作とを非同期に行う（ライトバック）ことにより，ライトペナルティの軽減をはかることが多い．

また，RAID-1とRAID-5との間でデータをダイナミックに割り振り，ライトペナルティの低減とディスク容量のバランスをとるHot Mirroring[8]という手法も研究されている．

3. 信頼性向上技術の実例

メインフレーム向けファイル装置では古くからさまざまな信頼性向上対策がなされてきたが，最近になってRAID技術を導入したファイル装置

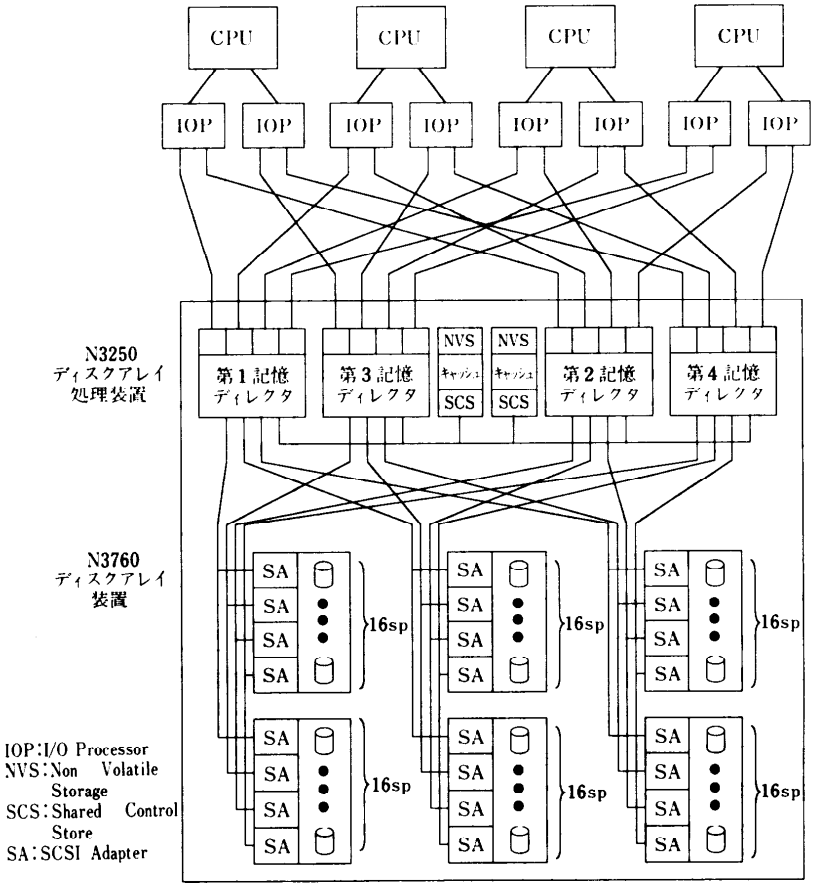

図-8 ファイル装置の構成例とホストとの接続例

が増えてきた[9]．本章では，実際のシステムで採用されている冗長化，故障診断，障害復旧技術について紹介する．また，PCやWSにおける信頼性向上の動向についても簡単に述べる．

3.1 冗長化

ファイル装置の構成要素としては，ディスクドライブ，コントローラ，コントローラ間バス，キャッシュメモリ，コントローラ－ホスト間接続，電源，配電ケーブル，ファン，などがある．

大型のファイル装置では，これらのほとんどを冗長化（たとえば$n+1$構成）している．**図-8**はNEC製ディスクアレイN 3250（ディスクアレイ処理装置)/N 3760（ディスクアレイ装置）のブロック図，および，ホストとの結合例である[10]．このディスクアレイはRAID-3を採用し，かつ同時アクセス可能なユニットを増やすことによりトランザクション処理にも適した構造をもつ．

ディレクタは，ディスクアレイのコントローラであり，ホストやディスクアレイと複数のチャネルで接続されている．このため，ディレクタやチャネルに障害が生じても，すべてのディスクに対してアクセスが可能となっている．また，不揮発性メモリ（NVS)/キャッシュメモリも2重化されており，片方のモジュールが故障してもデータは失われない．

ここではメインフレーム向けファイル装置の冗長化の例を示したが，小型ディスクドライブそのものにおいても，2重化したファイバチャネルインタフェース[11]を備えたり，2組のポートをもつSSA（Serial Storage Architecture）インタフェース[11]を採用するなど，アクセスパスを2重化する試みも始まっている．

3.2 故障診断/障害回避技術

ディスクドライブそのものに対する故障診断技術としては，パトロールリード/ライトなどがある．パトロールリード/ライトは主にディスクの媒体面のチェックを行うものであり，一定時間ごとに小領域を順番にリード/ライトしてチェックする．また，パトロールシークは，ヘッドが長時間同じ位置に留まって媒体に悪影響を与えるのを防ぐために，一定時間ごとにシーク動作を行わせるものであり，一種の障害回避策である．パトロールシークやパトロールリード/ライトを行っている間はディスクへアクセスできないため，パトロールの頻度は性能とのトレードオフによって決められる．

チェック機能としては，このほかにキャッシュのチェックや，コントローラ自身に対するセルフチェック，ファンや電源などの異常監視などがある．

積極的にチェックを行う方法以外にも，通常の動作をモニタリングすることでチェックする方法もある．たとえば，ディスクのリード/ライトエラーの回数またはリトライの回数のログをとり，エラーやリトライの量が一定回数以上になった場合にそのディスクの交換要求を通知する方法である．WSなどにディスクドライブを直接接続する場合はホスト側に監視ソフトを組み込むが，RAID装置ではRAIDコントローラ側で監視する．さらに最近では，ディスクドライブそのものに監視機能をもたせるものも出てきている．

3.3 障害復旧に対するアプローチ

大規模なRAID装置では，保守回線が保守サービス会社へ接続されて集中監視されるものもあり，ディスクの異常が検出されると保守要員がディスクの交換/復旧を行う．故障が起きてから復旧までは1日程度かかることも多い．確率的には少ないが，この間に別のディスクが故障した場合，ディスク内のデータが失われてしまう（データロスと呼ばれる)．その対策としてホットスペアと呼ばれる手法が使われている．これはRAID装置内に予備のディスクを接続しておき，1台のディスクに障害が起きたとき，自動的に予備のディスクに切り替えて障害復旧を行う方式である．

ホットスペアではディスクを余分に用意しておく必要があるためコスト増となるが，予備のディスクを通常時の運用ディスクの一部に用いて性能向上をはかる方法（パリティスペアリング，分散スペアリング）もある[4]．

故障したディスクを交換する際，システムを止めることが許されない場合が多い．このため，システムが稼働中でもディスクやコントローラの交換ができるように活線挿抜と呼ばれる手法がとられている．また，コントローラ内のソフトウェアに関しては，保守用の回線から無停止で保守およびバージョンアップ可能であるものも多い．

障害ディスクを交換すると，新しいディスク上に元のデータを再構築していく．RAID-3～6の

998

データの再構成は該当ディスク以外のディスクのデータのXORをとってデータを復元していくが，これは，障害ディスクのデータをホストから読み出すときの手順と同一であるため，障害ディスクのリード/ライト時にデータの復旧を兼ねてしまうことも多い．

3.4 PC/WSにおけるディスクの信頼性向上技術

以上みてきたように，メインフレーム用ファイル装置ではさまざまな信頼性向上対策がとられているが，PCやWSではコスト面から一部に限定せざるをえないのが現状である．このため，PCやWS用のRAID装置では，要求されるスループットや信頼性に応じて，1)ソフトウェアによりRAID機能を実現したもの，2)パリティはハードウェアで生成するがディスク—コントローラ間のチャネルを減らしてコストを下げたもの，3)スループットは高くなるように構成するが冗長度を少なくしたもの，などが使われている．

4. リアルタイム系アプリケーション用障害対策

ビデオオンデマンド（VOD）サービスや映像編集のようなリアルタイム系アプリケーションでは，データの可用性を高めるためにRAID装置を使うことが多い[12]．中でも中規模以上のVODでは，同時アクセス数をあげるために複数のディスクにまたがってデータが格納されている．このため，1つのディスクの障害がシステム全体の障害になることも多く，RAID装置に対する期待が大きい．しかしながら，従来のRAID装置は非リアルタイム系での利用を中心として考えられているため，シーク/リード/ライト時に生じるエラーに対してリトライの回数を大きく設定することにより，極力ディスク不良になる率を抑えている．リアルタイム系の記憶装置としては，障害時の応答性についても注意を払わなければならない．

また，一般にRAID-5を用いると縮退運転時にスループットが落ちるが，VODのようにスループットの低下が利用可能ユーザ数の減少に直結する場合，縮退運転時のスループットでシステム設計をせざるをえない．この観点から，VODサービスでは，縮退運転時でもスループットが落ちないRAID-3を採用するなどの考慮が必要になる．

5. おわりに

本稿では，RAID装置を中心にディスクアレイの概要と，実際のシステムにおけるファイル装置の信頼性向上の取組みについて概観した．また，VODサービスや映像編集における障害対策のあり方について簡単にまとめた．今後，ディスク装置の小型化と大容量化はますます進むものと考えられ，ディスクの信頼性向上技術の重要性は一層増すものと考えられる．一方，ディスクのインタフェースはファイバチャネルなどにみられるように，さらに高速化が進むものと期待される．ディスクの信頼性向上技術の研究開発もこの環境を前提に今後種々取り組まれるものと思われる．

参考文献

1) Patterson, D. A., Gibson, G. and Katz, R. H.: A Case for Redundant Arrays of Inexpensive Disks (RAID), Proc. of ACM SIGMOD, pp. 109-116 (1988).
2) Chen, P. M., Lee, E. D., Gibson, G. A., Katz, R. H. and Patterson, D. A.: RAID: High-Performance, Reliable Secondary Storage, ACM Computing Surveys, Vol. 26, No. 2, pp. 145-185 (1994).
3) Lee, E. D., Chen, P. M. and Gibson, G. A.: RAID-II: A High-Bandwidth Network File Server, IEEE Computer Architecture, pp. 234-244 (1994).
4) 喜連川：最近の二次記憶装置：ディスクアレイ，情報処理，Vol. 34, No. 5, pp. 642-651 (May 1993).
5) The RAID Book (Fifth Edition): The RAID Advisory Board (Feb. 1996).
6) Cao, P., Lim, S. B., Venkataraman, S. and Wilkes, J.: The TickerTAIP Parallel RAID Architecture, Proc. of the International Symposium on Computer Architecture, pp. 52-63 (May 1993).
7) 横田：データ再構築ネット（DR-net）における不均衡対策，信学技報，FTS 93-20, pp. 9-16 (1993).
8) 茂木，喜連川：Hot Mirroringを用いたディスクアレイのディスク故障時の性能評価，信学技法，CPSY 95-82, pp. 19-24 (1995).
9) 出揃った並列汎用機とディスクアレイ，日経ウオッチャーIBM版別冊，日経BP社（1995).
10) 橋本他：高速・大容量ディスクアレイサブシステム，NEC技報，Vol. 48, No. 9, pp. 65-70 (1995).
11) シリアルSCSIがいよいよ市場へ，日経エレク

トロニクス, No. 639, pp. 75-94(1995).
12) Misawa, K., Tsujisawa, T., Sugimoto, K., Kitamura, H., Shimoji, M. and Nakashima, S.: Disk I/O Scheduling and Communication I/O Schemes for Multimedia Sever HYPERMS, NEC Research & Development, Vol. 36, No. 3, pp. 417-428(1995).

(平成8年7月11日受付)

菊地 芳秀(正会員)

1961年生. 1986年東京工業大学理工学研究科修士課程修了. 同年NEC入社. 現在機能エレクトロニクス研究所メカトロニクス研究部主任. 検索システム, ディスクアレイ, VODシステムの研究開発に従事. 1993年スタンフォード大学計算機科学科訪問研究員. IEEE会員.

辻澤 隆彦(正会員)

1955年生. 1984年北海道大学工学研究科電気工学専攻博士課程修了. 同年NEC入社. 現在機能エレクトロニクス研究所メカトロニクス研究部課長. ディスク装置のI/O制御の研究に従事. 工学博士. 電子情報通信学会, システム情報制御学会, 計測自動制御学会各会員.

右田 守彦

1947年生. 1970年東京農工大学工学部電気工学科卒業. 同年NEC入社. 現在ファイル装置事業部第二システム技術部長. ファイルサブシステムの開発に従事. 電子情報通信学会会員.

赤木 正信(正会員)

1945年生. 1968年東京大学工学部電子工学科卒業. 同年NEC入社. 以来, 大型コンピュータ中央処理装置の開発, ディスクアレイ装置の開発に従事. 現在ファイル装置事業部事業部長代理. 電子情報通信学会会員.

1000