二重化内部データバスを持つＲＡＩＤシステム+

7Ｋ－3

新島秀人　宗藤誠治　村田浩樹　高橋伸彰
日本アイ・ビー・エム株式会社　東京基礎研究所

１．まえがき

近来、ＲＡＩＤといわれるディスクアレイシステムが脚光を浴びている。これは複数のハードディスクを並列に動作させてＩ／Ｏの高速化を図り、冗長構成により信頼性の向上を図ったものである．ＲＡＩＤ５はＳＣＳＩ－２の規格にあるコマンド並列処理（Tagged Command Queueing）の機能によりＲＡＩＤ３と同等程度の高い転送速度を実現する事が可能となるが、書き込み時のオーバーヘッドが大きく、高速なライトアクセスタイムを要求するような応用プログラムには適さない。これは偏にＲＡＩＤ５に於いては一つの書き込み要求に対して四回のＩ／Ｏを必要とする事が影響しており、この操作を見掛け上如何に減らすかがＲＡＩＤ５高速化への課題となっている。現在、この書き込み時間が実効的に見えなくなるようディスクキャッシュやライトアシストディスク（ＷＡＤ）を搭載することが検討され、実用化されているが、これらは何れもキャッシュやＷＡＤの容量を超える書き込みに対して無力であり、根本的なディスク本体への書き込み速度の改善が重要である事は論を待たない。

２．従来技術

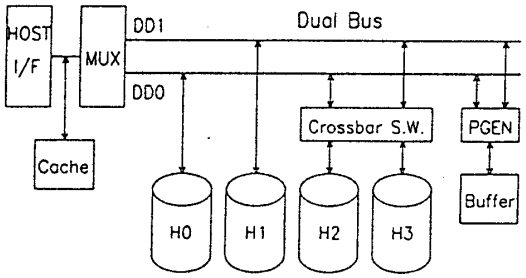
従来、小規模ネットワーク又は個人での使用を目的とした小型ＲＡＩＤシステムには、ソフトウエアＲＡＩＤを含め多くの物が出荷されている。ソフトウエアを用いた製品は最も安価で手軽にＲＡＩＤの信頼性を得る事が出来るが、パリティ生成のためにホストのＣＰＵリソースを必要とし、さらにデータ書き込み時には四回のデータ転送のためにホストのデータバスを占有してしまう。このため、ワークロードの重いサーバーに導入する場合にはサーバー全体のパフォーマンスに対する考慮が必要である。これに対し、ハードウエアで実現されたＲＡＩＤ製品群は高価とはなるが、専用のパリティ生成装置を装備し、また、書き込みの際にもホストは一度だけデータを転送すれば済むのでホストのＣＰＵやバス占有率に影響を与えることは無くて済む。ハードウエア製品の中でもＡＴＡドライブをシングルバスに接続した製品はハードウエア量が比較的少なくて済み、ハードウエアＲＡＩＤ製品群の中では最も安価な部類に入る。しかし、データ書き込みに際して四回のＩ／Ｏを順々に行わなければならず、ＲＡＩＤとしてのパフォーマンスはソフトウエアＲＡＩＤのそれと殆ど変わらない。ＳＣＳＩドライブを用いたＲＡＩＤ製品は比較的高価となるが、基本的には内部データバスをドライブ台数分持つため全てのドライブが並列動作が可能であり、このためデータの書き込み動作を

１）旧データ、旧パリティの読み出し
２）新データ、新パリティの書き込み

の２ステップで完了することが出来る。また、故障時やデータ回復処理時には故障ドライブ以外の全てのドライブから一度にデータを読み出し、全体の排他的論理和を計算することによりオリジナルデータを１ステップで回復出来る。

３．デュアル・データバス構成

ＡＴＡドライブを用いたシングル・データバス構成は安価であるがパフォーマンスに難点があり、ＳＣＳＩドライブを用いた構成では各ドライブに対して夫々一つのコントロールチップを必要とするため部品数が多く高価でかつ制御も複雑に成りがちである。そこで我々はＡＴＡドライブを用いてＳＣＳＩドライブを用いたのと同等程度の性能を持つ構成としてデュアル・データバスのＲＡＩＤシステムを製作した（図１）



MUX: Multiplexer
PGEN: Parity Generator

図１：デュアルバスＲＡＩＤシステムの構成

ＲＡＩＤシステム、特にＲＡＩＤ５の構成に於てはそのパリティーデータの生成及び書き込み動作を効率良く行う事が重要となるが、データの書込み動作では旧データ、旧パリティーとを読み出して同一のディスクにそれぞれ新データ、新パリティーとを

+ RAID System using Dual Data Bus Architecture
Hideto Niijima, Seiji Munetoh, Hiroki Murata, Nobuaki Takahashi
IBM Japan Ltd., Tokyo Research Laboratory. Shimotsuruma 1623-14, Yamato, Kanagawa 242

書き込まなければならない。即ち、どのような構成を取ろうとも一つのディスクに対して二回のＩ／Ｏが起こる事は避けられず、逆に言えばＲＡＩＤ５システムに対するデータ書込みの理論的な最少時間は２回分のディスクＩ／Ｏ時間である事が分かる。これはＳＣＳＩドライブを用いた構成がＲＡＩＤ５システムとしての最少時間で書込み動作を行っている事を示すが、この構成に於ても一時期にデータ転送を行っているドライブは高々二台であり、データバスが少なくとも二本あれば論理的な最少時間でデータの書込み動作を完了できる事を表している。以上の考察に基づき製作を行った物が本デュアル・データバス構成のシステムであり、

1）任意の組み合わせの二台のドライブに対し常に別々のデータバスを割り当てる事が出来る
2）ＡＴＡドライブを用い、データの転送は二台のドライブの用意が整ってから同時に一斉に行う
3）排他的論理和の計算はディスクからのデータ転送中にOn-the-flyで行う

を、基本設計方針としている。方針１）を満たすためには各ドライブからデュアル・データバスに対してデータセレクタを装備するのが設計上最も簡単であるが、実装上の観点からディスク二台に一つのクロスバー・スイッチを設け任意の組み合わせの接続が出来るようにした。また、各データバス一本について一台のドライブを固定接続しても方針１）を満たす事ができる。図１ではそれぞれのデータバスに一台ずつのディスクを固定接続してある。方針２）はＡＴＡドライブの特性を生かしたものである。ＡＴＡドライブは一度ＲＥＡＤＹ（ＤＲＥＱ−）を返してくれば必ず要求データの読みだし、書込み操作を一定のタイミングで完了する事が出来るが、ＳＣＳＩドライブはデータ転送をターゲット側が制御するため、二台のうち片方がディスコネクト要求を出すなどをしてタイミングが一定せず、複数ドライブからの同期転送が複雑になる。さらに、方針３）を実現する事によりパリティー計算時間を隠蔽し、ＲＡＩＤ５としての最少時間でのデータ書込み操作を実現できるようにした。

図１ではドライブを四台として示したが、これ以上の台数も当然対応可能である。その場合、追加ドライブ二台毎に対し一個のクロスバースイッチを用意してその出力をデュアル・データバスのそれぞれに接続すればよい。

４．動作説明

図１を用いてデータの書込み動作の説明を行う。今、書き込もうとしているデータの旧データがＨ０にあり、旧パリティーがＨ３にあるものとする。この時、新データ、新パリティーを書くべきディスクはそれぞれＨ０、Ｈ３である。この時、書込み動作を以下の順序で行われる。

1）Ｈ０、Ｈ３にデータ読だしコマンドを送る
2）Ｈ３をＤＤ１のデータバスに接続
3）両ドライブの用意ができたらデータ転送。ＰＧＥＮは両データの排他的論理和を計算し、結果をバッファに保持する
4）Ｈ０、Ｈ３にデータ書込みコマンドを送る
5）両ドライブの用意が出来たら、新データをＤＤ０に転送する。この時、ＰＧＥＮはＤＤ０のデータとバッファのデータとの排他的論理和を計算し、ＤＤ１に結果を転送する。Ｈ０とＨ３はそれぞれＤＤ０、ＤＤ１上のデータを書き込む

以上の操作で、実際のデータ転送時間として見えるのは３）と５）とであるので、見かけ上２回のＩ／Ｏでデータの書込みが終了する。

５．まとめ

二重化されたデータバス構成を持つＲＡＩＤシステムの構成とその動作とを説明した。本デュアル・データバス構成の特長として

1）多データバス構成のＲＡＩＤと比較して遜色の無い書き込みパフォーマンス
2）少ないハードウエア部品数
3）ＡＴＡドライブの特性の活用

等があげられる。本構成と他の構成との比較を表１に示す。但し、性能は計算による期待値である。

表１：ＲＡＩＤ構成の比較

ＲＡＩＤ構成	ソフトウエア	シングルバス	デュアルバス	多重バス
コスト	低	中	中	高
バス本数	1	1	2	5 -
部品数	無	小 - 中	小 - 中	大
ホスト負荷	有	無	無	無
書込み性能(単独)	1.0	1.0	1.5	1.5
書込み性能(複数)	1.0	1.0	1.66	1.66 - 2.5
読出し性能(複数)	1.0	1.0	1.3 - 1.5	1.5 - 1.8
故障時性能	1.0	1.0	1.2 - 1.3	1.5 - 1.7

６．参考文献

「ディスク・アレイ装置、性能向上で分散システムの要に」日経エレクトロニクス,1993.4.26(no.579) pp.78-103