5G－9

XOR付きクロスバースィッチを用いた Dual-Bus RAID システム

宗藤 誠治、 村田 浩樹
日本アイ・ビー・エム(株) 東京基礎研究所

1 はじめに

Dual-Bus RAID システムは内部バスを2重化する事により、RAID level 4,5 の書き込み要求処理を効率よく行なう。本稿では内部バスとディスクを接続するクロスバースッチ内にＸＯＲ演算機能を付加する事により、本方式で問題であった障害発生時のデータの読みだし、及びデータの復旧を高速化する手法について報告する。

RAIDの性能を測る上で重要な要素として、ディスククラッシュなどの障害発生時での読みだし、書き込み速度、及びドライブの復旧時での障害回復時間がある。Dual-Bus RAID システムでは、障害時におけるこれら複数台(3台以上)のドライブ間でのパリティ演算を必要とする場合に、同時に出来るXOR演算数に限りがあるため、複数の段階に分けて各ドライブ、パリティバッファー間のデータ転送、XOR演算を行なう必要がある。その結果、ストライプを構成するドライブ数と同数のバスを持ち、同時に複数ドライブ間でのXOR演算が可能なシステムに対して実際のデータ転送時間に関して性能的に劣る事になる。

この問題の解決策の1つとして、ドライブ2台と共有バス2本とを相互接続するクロスバースィッチ内部に排他的論理和(XOR)回路を設ける。その結果、任意のチャンネル間でのXOR演算を可能となり、適切なデータパスの設定により効率よくXOR演算を行なう事が出来る。また、データ復旧速度が高速になる事により、システムのMTBFがさらに向上する。

2 H/W構成

図1にＨＤＤ数4台で構成した場合のシステムのブロック図を示す。各ＨＤＤはクロスバースィッチを介して2本の共有バスに接続される。通常の動作時には、読みだしでは該当するドライブが1台どちらかの共有バスに接続され、読みだしデータはそのままホストに転送される。書き込み時では、RAID5などのパリティ更新を伴う場合、該当するデータドライブ、パリティドライブに対してRead-Modify-Write操作が2本の共有バスを介して効率よく2段階のデータ転送により実現される。

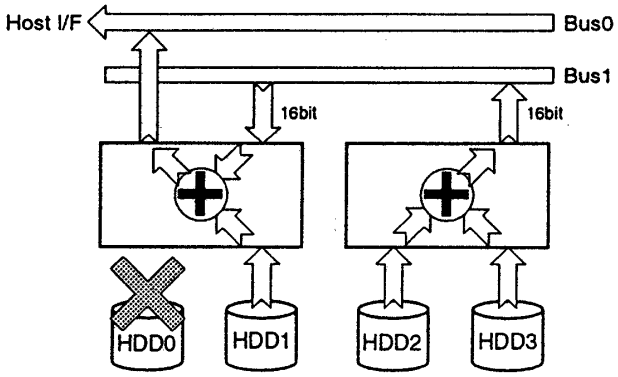

図2 故障ドライブのデータ読みだし操作

今回、新たにＸＯＲ演算機能をドライブと共有バスとを接続するクロスバースィッチ内に追加する。その結果、データパスの経路内で2もしくは3個のデータ間でのXOR演算が可能となり、次に示すように、複数データ間のXOR演算の効率が向上する。

3 データ転送

4台構成(データドライブ3台、パリティドライブ1台)でのHDD0が故障したケースについて、その故障ドライブに対するデータ読み出し要求時のデータの流れ図2に示す。

故障したHDD0に対する読みだし要求があった場合、そのデータはストライプを構成する残り3台のデータのXOR演算から求められる。まずHDD2,HDD3から読みだしたデータをクロスバースィッチ内のXOR演算回路を用い HDD2⊕HDD3 の演算結果に変換してBus1に出力する。同時にBus1に出されたデータとHDD1からのデータをもう一つのクロスバースィッチ内のXOR演算回路を用いて演算し、最終的にBus0には HDD1⊕HDD2⊕HDD3=HDD0 のデータが出力される事になり、ホストに対し転送される。以上のデータ転送はOn-the-flyでディスクアクセスと同期して行われ、単体のドライブと同等のデータ転送レートが得られる。

同様に、データの復旧の場合は復旧したHDD0に対

Dual-Bus RAID system using Crossbar Switch with XOR
Seiji Munetoh, Hiroki Murata
IBM Research, Tokyo Research Laboratory

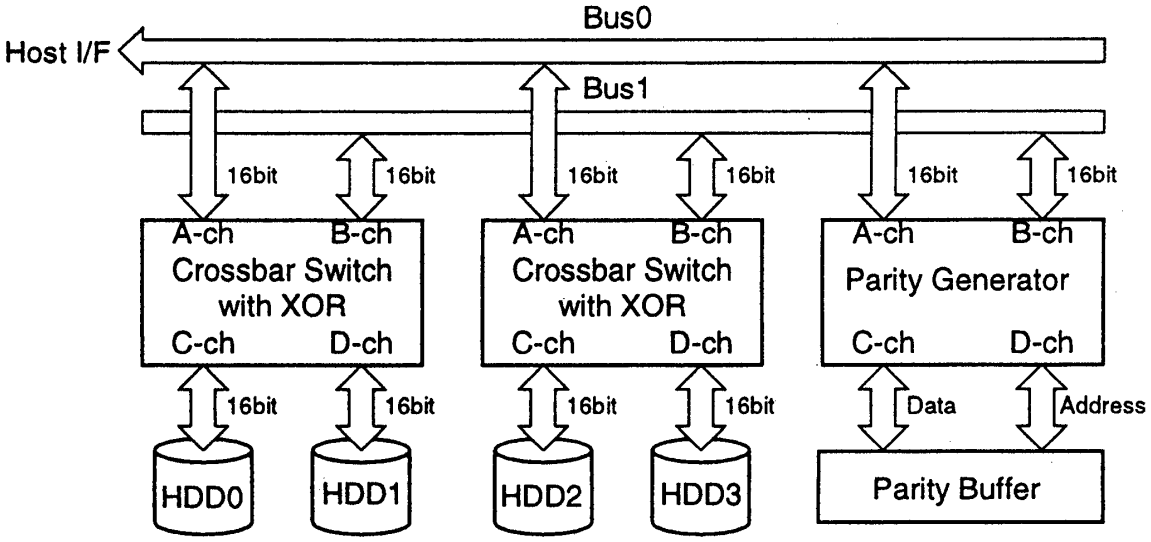

図1 ＨＤＤ４台でのシステム構成

動作		従来	XOR付き
通常	読みだし	1 step	1 Step
	書き込み	2 step	1,2 Step
障害	読みだし	2 step	1 Step
	書き込み	2 step	1,2 Step
復旧		2 step	1 step

表1 従来との比較(ドライブ数4台)

して再生したデータを書き戻せば良い。つまり、読みだしではBus0に出力していたHDD1⊕HDD2⊕HDD3をHDD0に対し書き込みとしてデータ転送を行なう事により完了する。

表1にHDD数4台におけるRAID 5システムでの従来との転送処理の比較を示す。最も単純なデータ転送要求パターンにおける、データ転送のステップ数で比較する。これはデータ転送を行なっている時間だけでの比較の為、ディスクのシーク、回転待ち時間、コマンドオーバーヘッドなどを含めた場合でのトータルな性能を示すわけではないが、データ転送に関する限りにおいては性能向上が期待できる。また、複数データのXOR演算可能なシステムに対しても遜色無いオペレーションが可能である事がわかる。

4 まとめ

Dual-Bus構成のRAIDシステムにおいて、クロスバースッチにＸＯＲ演算機能を追加する事により、RAIDの各種オペレーションが効率良く実現できる事を示した。

大きなデータの連続転送を考える場合、ホストI/Fの転送能力を十分に生かそうとすると、ドライブのメディア・トランスファー・レートが律速になる。その為、十分な数のディスクを用いてアレイを構成する必要がある。

ディスクの数が4台までの小規模なRAIDシステムでは今回の機能拡張は有効であるが、5台以上に拡大した場合、共有バスが2本であるために複数回に分割した転送、XOR演算が必要になり、障害発生時の性能低下が懸念される。

参考文献

Paterson,D.A, et al,"A case for redundant arrays of inexpensive disks (RAID)",ACM SIGMOD 88,pp. 109-116 (Jun. 1988)

新島, 他: "二重化内部データバスを持つRAIDシステム", 第49回情処全大,7K-3

宗藤, 他: "ATAドライブを用いたRAIDサブシステムの構成", 1994年信学秋期全大,D-88