大規模並列ディスクの結合形態と性能の評価

6F－2

味松康行　横田治夫

北陸先端科学技術大学院大学 情報科学研究科

1　はじめに

Patterson らが提案した Redundant Arrays of Inexpensive Disks(RAID) は、並列に動作可能な複数のディスクを１つのシステムに統合することで性能の向上を図り、またシステム内に冗長な情報を持つことでシステム全体の信頼性低下を防ぐ [1]。しかし、大規模な RAID システムではディスクを結合するバスや単一のインタフェースがボトルネックとなり、ディスク台数に見合う性能が得られないことが考えられる。

この問題を解決するために、我々は Data Reconstruction networks(DR-net) を提案してきた [2][3]。DR-net は RAID と異なり、ディスクをバスではなくネットワークで結合する。これによりバス上での通信ボトルネックを解消することができる。また、独立に動作する任意個のインタフェースを持つことができるため、インタフェースでの負荷の集中を緩和することができる。また、新たな冗長情報の利用を導入することで信頼性が向上している [4]。

本稿では、シミュレーション実験を通して DR-net とバスを用いるシステムを比較、評価することにより、ディスクをネットワークで結合し複数のインタフェースを用いる DR-net の構成の有効性を示す。

2　DR-net の概要

DR-net はディスクノード、インタフェースノード、およびそれらを結合するネットワークからなる。ディスクノードにはディスクおよび XOR 演算器があり、インタフェースノードには外部との通信インタフェースがある。冗長情報の管理単位であるパリティグループは、データを保持する複数のディスクノード (データノード) と、それらのデータのパリティを保持する１つのディスクノード (パリティノード) から構成される。データの更新にはパリティの更新が伴い、新しいパリティは古いパリティ、新しいデータ、古いデータの排他的論理和により計算される。従って、更新のためにはディスクの書き込みの他に古いデータ、パリティの読み出しとノード間通信が必要となる。

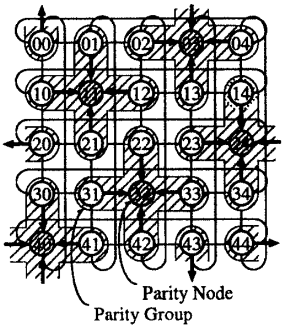

図 1: FPG(First Parity Groups) の構成

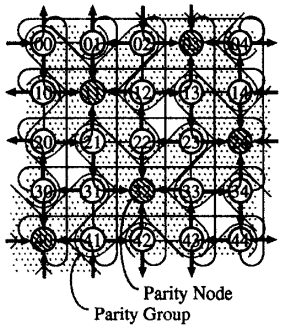

図 2: SPG(Second Parity Groups) の構成

DR-net では 2 種類のパリティグループを用い、それぞれ First Parity Groups(FPG), Second Parity Groups(SPG) と呼ばれる。5 × 5 のトーラスネットワークを用いる DR-net における FPG, SPG の構成を図 1, 2に示す。これら 2 つのパリティグループを併用することにより、任意の 2 つのディスク故障が発生してもデータを再構築することができる。

DR-net のインタフェースノードは、ネットワークのノード間リンク部に設置される。インタフェースノードはネットワーク内の任意の位置に置くことができる。また、インタフェースノードの数に制限はなく、インタフェースでの負荷の集中を緩和することができる。

3　シミュレーション実験

実験では DR-net およびバスを用いるシステムについて、ディスクアクセス要求のレスポンスタイムとスループットを測定する。比較対象のバスシステムは、すべてのディスクを１本のバスでつなぐフラットなシステム、FPG 毎にクラスタ化した階層化バスを用いるシステム、FPG, SPG の 2 段階でクラスタ化した 3 レベルの階層化バスシステムである。

実験では、ランダムに選んだディスクブロックに対

Performance Evaluation of a Large Disk System with an Interconnection Network

Yasuyuki Mimatsu, Haruo Yokota
School of Information Science, Japan Advanced Institute of Science and Technology

表 1: シミュレーションに用いた主なパラメータ

バスバンド幅 (上層)	1000 Mbits/sec
バスバンド幅 (中層)	100 Mbits/sec
バスバンド幅 (下層)	40 Mbits/sec
DR-net リンクバンド幅	40 Mbits/sec
通信セットアップ時間	10 μ sec
セクタサイズ	512 bytes
ブロックサイズ	32 セクタ
平均シーク時間	19 msec
ディスク転送速度	2.66 Mbytes/sec

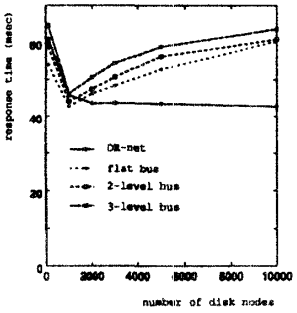

図 3: 読み出しのレスポンスタイム (インタフェース数=40)

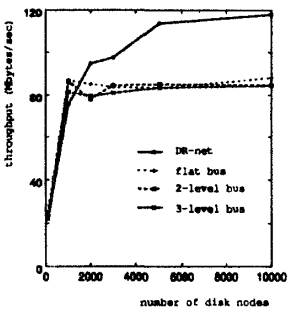

図 4: 読み出しのスループット (インタフェース数=40)

してインタフェースノードからアクセス要求を発行する。各ノードは一度に 1 つの要求を処理する。要求を処理中のノードに対する要求は、処理が終るまでインタフェースノードで待たされる。インタフェースノードは、発行を待つ必要がない限り、次々に要求を発行する。シミュレーションに用いた主なパラメータを表 1に示す。

図 3, 4は、ディスク数を 10000 まで変化させたときの読み出し要求に対するレスポンスタイムおよびスループットである。インタフェースノード数は一定 (40) である。ディスク数が増えるに従い、バスを用いたシステムでは階層化のレベルにかかわらず、レスポンスタイムは増大し、スループットも頭打ちになることがわかる。一方、DR-net ではディスク数が増えてもレスポンスタイムはあまり変化せずほぼ一定に保たれている。また、スループットもバスを用いたシステムよりも高く、ネットワークを用いてバスの通信ボトルネックの影響を解消した効果が示されている。

図 5, 6は、インタフェース数を 120 まで変化させたときの読み出し要求に対する結果である。ディスクノード数は 10000 で一定である。インタフェース数の増加により多くの要求が並行して処理されることになるが、DR-net ではレスポンスタイムの増加は緩やかで、しかもスループットが大きく向上している。バスシステムでは並行して処理する要求の増加により、バスのコンテンションが激しくなるため性能はあまり向上しない。

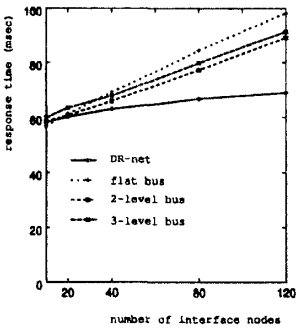

図 5: 書き込みのレスポンスタイム (ディスク数=10000)

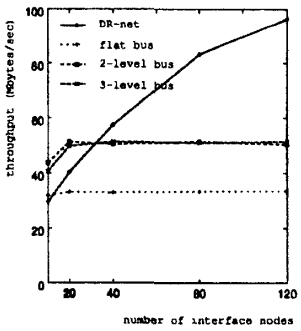

図 6: 書き込みのスループット (ディスク数=10000)

4 おわりに

本稿では、大規模な構成のディスクシステムにおいてディスクの結合形態が性能に与える影響について述べた。従来のバスを用いるシステムでは、インタフェース- ディスク間のすべての通信がバスを通るため、バスがボトルネックとなりディスク台数やインタフェース数を増やしても性能は向上しない。我々が提案している DR-net はバスではなくネットワークでディスクを結合するため、通信ボトルネックは存在せず、ディスク台数に応じた性能が得られる。シミュレーションで用いた DR-net の個々のリンクスピードは、最も遅いバスのスピードと同程度であるにもかかわらず、システム全体では高い性能が得られることが示された。今後、ディスクの単体性能が向上した場合、通信の処理時間に占める割合は増加し、バスのボトルネックの問題はますます重要となる。高性能かつ距離の長いバスを作ることが難しいことを考えると、ネットワークを用いる DR-net は大規模構成に適したシステムであると考えられる。

参考文献

[1] D. A. Patterson, G. Gibson and R. H. Katz: "A Case for Redundant Arrays of Inexpensive Disks(RAID)", Proc. of ACM SIGMOD Conference, pp. 109–116 (1988).

[2] 横田治夫: "RAID のネットワーク上への展開と信頼性向上", 信学技法 CPSY 93–11, 電子情報通信学会 (1993).

[3] 味松康行, 横田治夫: "ネットワーク結合並列ディスクにおける耐故障制御の影響", 情報処理学会論文誌, 37, 7, pp. 1419 – 1428 (1996).

[4] 味松康行, 横田治夫: "並列ディスクシステムのパリティグループの構成の変化と信頼性の比較", 信学技報 FTS 95–34, 電子情報通信学会 (1995).