4K－5

# 階層構成を用いたディスクアレイにおける記憶効率の向上法に関する一考察

茂木 和彦　喜連川 優
東京大学 生産技術研究所

## 1 はじめに

2次記憶装置の高性能化・高信頼化を目的とした冗長情報を記録するディスクアレイ (RAID)[1] の開発が進められている。その中で、サイズは小さいが多数のアクセス要求があるような負荷では、ミラー (RAID Level 1) や RAID5 (RAID Level 5) が良いと考えられている。RAID5 ではパリティを用いた冗長化を行っており、データ書き込み時のパリティ更新ためのオーバヘッドやディスク故障時のデータ復旧作業の影響による性能の低下が問題となっている。この点に関して優れているミラーでは、データのコピーを保持することによる冗長化を行っており、データ容量が少ないという問題点が存在する。これらの問題を解決するために、Hot mirroring と名付けた記憶管理法を提案した [2]。本方式は、ミラーと RAID5 を階層的に用いることにより高い性能と高い記憶効率を目指すものである。本稿では、Hot mirroring を用いたディスクアレイのアクセスコストと、ミラー領域の割り付け量の動的削減時の性能予測法について検討する。

## 2 Hot mirroring のアクセスコスト

Hot mirroring では各ディスクをそれぞれ2つの領域に分割し、アクセスローカリティを利用してアクセス頻度が高いものはミラー、低いものは RAID5 と同様な方式で記録し、双方の特性をうまく引出すことにより高性能化・大容量化を図る。Hot mirroring においては書き込みは全てミラー領域に対して行われる。しかし、ミラー領域の容量には制限があるので、アクセス頻度が低い (コールドな) ものを適宜 RAID5 領域に書き戻す必要がある。(以下、この動作をマイグレーションと呼ぶ。) ホット領域の空きブロック数がその下限値を超えた時、最終アクセスからの経過時間が最も長いものをコールドブロックと見做し、そのブロックをコールド領域へマイグレートする。(LRU アルゴリズムを用いる。)

マイグレーションは移動されるデータの読み出しと RAID5 領域への書き込みを必要とする。従って、$w$ を書き込み率、$m$ をデータ書き込み時のマイグレーション実行率、$t_{acc}$ を1ブロックのアクセス実行時間、$t_{rmw}$ を1ブロックのリード・モディファイ・ライト アクセスの実行時間とし、先着順でディスクアクセスを実行した時、Hot mirroring を用いたディスクアレイの1アクセス当たりの平均ディスク占有時間は以下のように近似可能である。

$$(1+w)t_{acc} + mw(t_{acc} + 2t_{rmw}) \quad (1)$$

## 3 ミラー領域の割り付け量の動的削減と性能予測

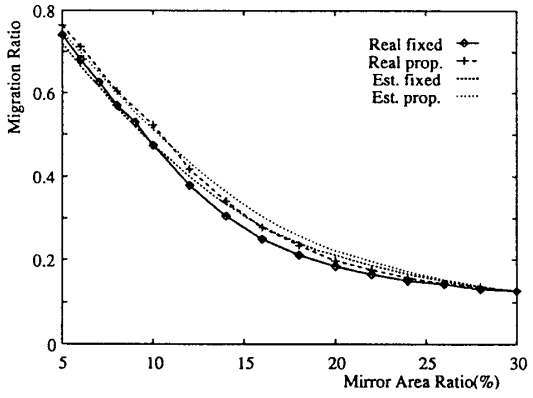
Hot mirroring において、データの記憶容量を増やすためにはミラー領域への割り付け量を減らす必要がある。しかし、その量を減らしすぎると、今度はマイグレーションが頻繁に実行されることになり、大きな性能低下が引き起こされる。Hot mirroring を効率的に利用する時、与えられた負荷とミラー領域への割り付け量からどの程度の性能が得られるかを見積もり、ミラー領域をどの程度まで削減可能か判断する必要がある。

アクセスコストを示す式 (1) において、$t_{acc}$ と $t_{rmw}$ は利用するディスクにより決定されるものである。また、与えられた負荷の特徴には大きな変化が無いことを仮定した時、$w$ もアクセスの統計を取ることにより容易に推定可能である。従って、マイグレーション実行率 $m$ が推定されれば、アクセスコストを見積もることができ、おおよその性能を推定することができる。

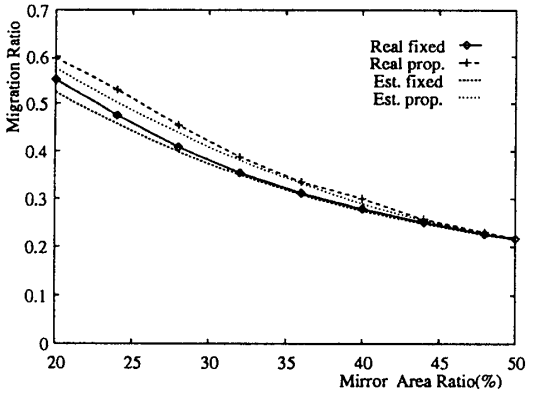
### 3.1 マイグレーション実行率の推定

マイグレーションは RAID5 領域に存在するブロックの更新によるミラー領域の消費が原因で引き起こされる。従って、RAID5 領域のブロック更新率とマイグレーション実行率はほぼ等価である。ディスクコントローラ上で、マイグレーションのターゲット選択の実装法として LRU リストを用いることが可能である。ミラー領域の削減は LRU リストのエントリ数が減少することを意味する。実際には LRU リストのエントリ数が変化するとそれぞれに対応する部分の振舞は必ずしも一致しないが、おおまかな傾向は一致することが期待できる。そこで、実際に管理に用いている LRU リストを用いて、ミラー削減時の書き込み動作の振舞を推定することを考える。ミラー領域内のブロックの更新時に LRU

A Study of Improving Storage Efficiency on Hierarchcal Disk Arrays
Kazuhiko Mogi and Masaru Kitsuregawa
Institute of Industrial Science, University of Tokyo




(a) 90-10




(b) 80-20

図 1: マイグレーション実行率の推定

リスト内の位置を調べ、その分布を測定する。このうち、エントリ数が削減された時にLRUリスト内に存在できない位置ものが、ミラー領域削減時にRAID5領域のブロック更新を必要とする(マイグレーションを引き起こす)ものであると推定することができる。

上述の推定法を評価するためにシミュレーションを行った。24台のディスクを用い、400 IOs/sec、書き込み率が30%であることを仮定する。90%のアクセスが10%の負荷に集中(90-10)したときに、ミラー領域に全ディスク容量の30%を割り当てた時のブロック更新時のLRUリスト内の位置分布から推定したマイグレーション実行率(Est.)とその実測値(Real)を図1(a)に、80%のアクセスが20%の負荷に集中(80-20)したときに、ミラー領域に50%を割り当てた時の分布からの推定値と実測値を図1(b)に示す。fixedは利用されるデータブロック数が変化しない場合を、prop.はミラー領域削減による記憶容量の増加分は全てデータブロックとして利用した場合を示す。prop.の推定においては、ミラーブロック更新時のLRUリスト内の位置はデータブロックの増分に比例して伸びたと仮定した。図のように、ミラーブロック更新時のLRUリストの位置分布から求めた値はマイグレーション実行率に近い値を示す。

| ミラー領域 | 10% | 20% | 30% |
|---|---|---|---|
| 実測値 (IOs/sec) | 600 | 850 | 850 (900) |
| 推定値 (IOs/sec) | 560 | 690 | 750 |
| 修正推定値 (IOs/sec) | 640 | 840 | 910 |

表 1: 最大許容負荷の推定値 (90-10, R:W=7:3)

### 3.2 最大許容負荷の推定

マイグレーション実行率が求まると、式(1)を用いて1アクセス当たりのディスク占有時間が求まり、それを用いて最大許容負荷を求めることができる。90-10アクセスローカリティで書き込み率が30%の時の、ミラー領域の割合を10%, 20%, 30%とした時の最大許容負荷の実測値(50 IOs/sec刻み)と、上述のprop.のマイグレーション実行率の推定値を用いて計算した推定最大許容負荷を表1に示す。なお、ディスクバンド幅の99%を利用可能として計算した。ミラー領域が30%の場合、ディスクバンド幅は900 IOs/secまで許容可能であったが、その負荷ではレスポンスタイムの大幅な悪化が起こり、実用的な許容負荷は850 IOs/secであった。

誤差の主な原因は、ミラー領域へのアクセス集中の効果により、式(1)中の$t_{acc}$に誤差が生じているためである。そこで、ミラー領域へのアクセスの平均シーク距離をミラーシリンダ数の1/3、RAID5領域への平均シーク距離を全シリンダ数の1/2と仮定し、ミラー領域とRAID5領域へのアクセス時間を分離する。この条件でデータ読み込み時のミラー領域へのアクセス率を(1-$m$)と仮定した時の修正推定値も合わせて表に示す。

## 4 まとめ

本稿では、Hot mirroringを用いたディスクアレイのアクセスコストとミラー領域の動的削減時の性能予測法について述べた。この性能予測を基にどの程度までデータ容量を増やすことができるかを決定することができる。今回用いた負荷は、LRUリストを用いたマイグレーション実行率の推定法に適した負荷であると考えられ、本方式の評価はまだ不十分であると言える。今後、種々の負荷を用いて妥当性の評価をより詳細に行う必要がある。

## 参考文献

[1] David A. Patterson, Garth Gibson, and Randy H. Katz. A Case for Redundant Arrays of Inexpensive Disks (RAID). In *Proc. of ACM SIGMOD*, pp. 109–116, June 1988.

[2] 茂木和彦, 喜連川優. Hot mirroringを用いたディスクアレイの基本性能評価. 情報処理学会データベースシステム研究会 95–DBS–104–7, 1995年7月.