# 階層型 RAID を用いた大規模仮想ディスク修復に関する考察

チャイ エリアント† 上原 稔† 森 秀樹†

東洋大学†

## 1 はじめに

近年ストレージサービスに対する要求は HDD 技術の進歩を促した。その結果、安価で大容量なディスクが入手可能となった。しかし、アプライアンス系のファイルサーバで使用している HDD はその普及品より 10 倍高価である。このように現在のストレージのコストは適切でない。ストレージが高価な理由は専用ハードウェアにある。普及品のハードウェアと専用ソフトウェアによって問題を解決できる。

我々は、大規模ストレージを構築するためのツールキット VLSD[2] (Virtual Large-Scale Disk)を開発した。VLSD は 100 % pure Java であるため、プラットフォームに依存しない。我々は VLSD を用いて 500 台の PC からそれぞれ空き容量 170GB を集め 70TB のストレージを構築するシステムを試作した。このシステムでは RAID66(2 階層の RAID6)を用いて十分な MTTF を実現している。

ストレージ全体の信頼性は MTTR にも依存する。ディスクが多くなるほど容量は大きくなるが、 MTTF は減少する。それゆえ MTTR を最小化することが必要である。MTTR を最小化するには、故障ディスクをスペアに交換する時間（交換時間）とデータを修復する時間（修復時間）をそれぞれ短縮する必要がある。交換時間はホットスペアが存在するとき無視できる。

本論文では、VLSD を用いた試作ストレージにおいて修復時間を測定し、システム全体の信頼性を評価する。

## 2 関連研究

RAID の信頼性ではシステムが故障するまでの平均時間（MTTF）の計算がよく用いられる。文献 [1]によると、RAID1 の MTTF は以下の式で表せる。

$$\frac{MTTF_{disk}^2}{N \times MTTR_{disk}}$$

RAID1 では N=2 である。RAID1 を一般化すると以下のようになる。

$$\frac{MTTF_{disk}^N}{N! \times MTTR_{disk}^{N-1}}$$

RAID4、RAID5 の MTTF は以下のようになる。

$$\frac{MTTF_{disk}^2}{N \times (G-1) \times MTTR_{disk}}$$

$MTTF_{disk}$ は一台ディスクの MTTF、N はディスクアレイのディスク数、G はパリティディスク含まないグループ内のデータディスク数、 $MTTR_{disk}$ は一台ディスクの MTTR（故障したから復旧にかかる時間の平均）である。RAID6 の MTTF は以下のようになる。

$$\frac{MTTF_{disk}^3}{N \times (G-1) \times (G-2) \times MTTR_{disk}^2}$$

シングル RAID を多重にすること（階層型 RAID）で性能や信頼性を上げることができる。例えば RAID0 を上位層にすると、性能が上がる。そして、下位層に RAID1、RAID5、または RAID6 にすると信頼性が高められる。本研究では信頼性を重視するために RAID55 や RAID66 が用いられる。RAID55 の MTTF は以下の式で表せる。

$$\frac{MTTF_{RAID5\_disk}^2}{N \times (G-1) \times MTTR_{RAID5\_disk}}$$

RAID66 の MTTF は以下の式で表せる。

$$\frac{MTTF_{RAID6\_disk}^3}{N \times (G-1) \times (G-2) \times MTTR_{RAID6\_disk}^2}$$

$MTTF_{RAID5\_disk}$ は RAID5 ディスクの MTTF で、$MTTF_{RAID6\_disk}$ は RAID6 ディスクの MTTF である。

信頼性には MTTR の影響も大きい。MTTR が大きくなると MTTF が小さくなる。ゆえに MTTF を大きくするため、つまりシステムが故障するまでの平均時間を大きくするのに MTTR を小さくする必要がある。

## 3 修復

冗長な RAID では、故障が発生しても動作し続けることができる。RAID5 は 1 台のディスクに耐える。RAID6 は 2 台のディスク故障に耐える。RAID から冗長性が失われた状態を縮退モードという。しかし、故障時の性能は正常時より低下する可能性がある。我々は文献 [3]にて故障時の VLSD の性能を評価した。その結果、速度低下はあるものの使用可能なレベルであることが確認できた。

2 節で述べたように RAID の MTTF はディスクの MTTF と MTTR に依存する。我々の大規模ストレージは階層型 RAID に基づくため、同様にその MTTF はディスクの MTTF と MTTR に依存する。

MTTR は、故障してから再び正常に使用できるまでの時間として定義される。縮退モードでは、RAID は停止しないが、もう 1 台故障すると停止する。RAID を稼動し続けるためには、もう 1 台故障する前に故障したディスクを修復する必要がある。

修復は以下のように行われる。故障したディスクをスペアディスクに入れ換えた後、RAID 内で故障したディスクの内容だけを読み取り、単 純に書き戻す。これにより、無故障ディスクからデータを復元し、そのデータを故障したディスクに書き込むことができる。VLSD では RAID クラスが故障ディスクに対応するブロックを順次アクセスすることで復元する。

## 4 評価

ここでは、修復時間を測定し、RAID 全体の信頼性を評価する。

A Case Study on Recovery of Large-Scale Virtual Disk using Hierarchical RAID

†Erianto Chai, Minoru Uehara, Hideki Mori - Toyo University

### 4.1 実験環境

本評価は、AMD Athlon(tm) 64 X2 Dual Core 3800+、メモリ 2GB、Windows XP Professional x64 で行われた。

### 4.2 HDD の MTTF

目標とする大規模ストレージでは、対費用効果 の高い 320GB の HDD を使用する。表 1 に代表的な製品の MTTF を示す。その平均は 85 万時間である。よって、平均的な HDD を 500 台使用した場合、85 万/500=1700[h]≒70[日]に 1 台の割合で故障する可能性がある。

表 1 HDD 製品の MTTF

| メーカー | 製品 | MTTF[$10^3$h] |
|---|---|---|
| Seagate | ST3320620AS | 700 |
| IBM | HDT725032VLA360 | 1000 |
| 平均 | | 850 |

### 4.3 修復時間の評価

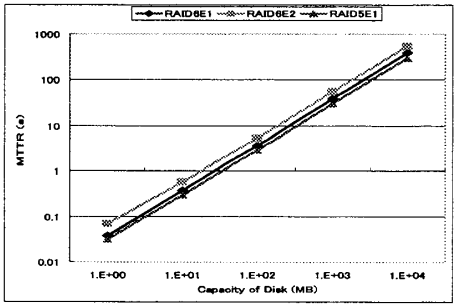


図 3 RAID5 と RAID6 の修復時間

ここでは、RAID5 と RAID6 の修復時間を評価した結果について述べる。 4 台の仮想ディスクで構築された RAID5 と RAID6 の修復時間を図 3 に示す。

RAID5E1、RAID6E1 はディスクが 1 台故障時の RAID5、RAID6 で、RAID6E2 はディスクが 2 台故障時の RAID6 である。グラフに示すように RAID6E1、RAID6E2 と RAID5E1 のディスク容量は MTTR に比例する。修復時間は RAID6E2 > RAID6E1 > RAID5E1 の順になる。

### 4.4 並列修復による MTTR

表 2 並列修復の評価

| N | 処理時間[h] |
|---|---|
| 1 | 8.47 |
| 2 | 7.12 |
| 4 | 7.97 |

修復作業は並列処理できる。もし N 並列で修復する場合、ブロック番号を mod N で分類か、あるいは 1/N で分類すればよい。

N が必要以上に大きくても効果は期待できない。 N=1、2、4 の場合、表 2 のような結果が得られた。この実験ではデュアルコアの場合、スレッド数が 2 のときが最適である。これは CPU より I/O がボトルネックとなっているためと考えられる。

### 4.5 信頼性の評価

表 3 2 階層 RAID の MTTF と容量効率

| RAID level | MTTF[h] | 容量効率[%] |
|---|---|---|
| RAID11 | $4.25 \times 10^{2463}$ | 0.2 |
| RAID15 | $1.7 \times 10^{219}$ | 4.3 |
| RAID16 | $2.2 \times 10^{328}$ | 4.1 |
| RAID51 | $4.19 \times 10^{182}$ | 4.3 |
| RAID55 | $5 \times 10^{15}$ | 91.1 |
| RAID56 | $1.1 \times 10^{23}$ | 86.8 |
| RAID61 | $4.78 \times 10^{272}$ | 4.1 |
| RAID65 | $7.8 \times 10^{23}$ | 86.8 |
| RAID66 | $2.2 \times 10^{35}$ | 82.6 |

測定した MTTF、MTTR を用いて階層 RAID 全体の MTTF を計算する。RAID1, 5, 6 の組み合わせからなる階層 RAID を考える。

ここで、ディスク総数を 484、下位層の台数を 22、上位層の台数を 22 にしたときの結果を表 3 に示す。なお、RAID3, 4 は RAID5 と等しいため、省略する。また、RAID1 は N 台に同じ内容をコピーするものとした。

RAID1 を用いると容量効率は極端に低下するので RAID1 を使うべきではない。容量が大きくて MTTF が高いのは RAID66 が最もよいが、RAID6 は普及したばかりで比較的高価である。

## 5 まとめ

本論文では、VLSD の修復機能と性能について述べた。VLSD は PC の空き容量を用いて大容量ストレージを構築するツールキットである。今回の評価で、負荷のない条件下では単一の 170GB ディスクを 7.12 時間で修復可能であることがわかった。これは 500 台の PC からなる RAID55 の 75TB 試作システムにおける MTBF が最悪 $5 \times 10^{15}$ 時間であることから、十分実用可能であるといえる。

今後はディスク交換における割り当てポリシーの評価を行う必要がある。例えば、実行時の性能を優先するなら局所ホットスペアがよい。あるいは、交換時間の 短縮を優先するなら大局ホットスペアがよい。それぞれの性能を評価する必要がある。

今回、1 台のマシンで評価した。しかし、実際のシステムは複数マシンで構成される。複数のマシンでネットワーク経由で使用した場合、性能はどう変化するのか調査する必要がある。

## 参考文献


[1] David A. Patterson, Garth Gibson, and Randy H. Katz: "A Case for Redundant Arrays of Inexpensive Disks (RAID)", ACM SIGMOD, 1988

[2] Chai Erianto, Minoru Uehara, Hideki Mori, Nobuyoshi Sato: "Virtual Large-Scale Disk System for PC-Room", LNCS 4658, Network-Based Information Systems, pp.476-485, (2007.9.3-4)

[3] Chai Erianto, Minoru Uehara, Hideki Mori: "Performance Evaluation at Failure in a Large-Scale Virtual Disk", DPSWS2007 (2007.10)