3耐故障RAIDの実装

中村　祐司†　　上原　稔†
東洋大学総合情報学部総合情報学科†

1 はじめに

今日、ブログなどのCGMの普及、写真ないし動画投稿サイトの流行などにより、かってないほどの情報がネットワーク上に氾濫している。Googleはこれらの情報をインデックス化しようと努力しており、ある程度成功しているように見えるが、まだまだ未分類の情報が実世界に放置されている。最近では、日々の生活を記録する人々が増えている。これはライフログと呼ばれる。ライフログで動画が使われるようになると情報の量はいっそう増大する。仮想世界は実世界の情報をすべて取り込もうとしている。仮想世界が実世界を飲み込もうとするとき、それらの情報を保存するストレージが必要になる。今日では、大量生産される安価なPCを用いて大規模なストレージを実現可能となった。また、HDD技術の進歩によりストレージ容量も増えている。

大規模ストレージでは多量のディスクを運用する必要がある。しかし、ストレージのディスク数が増加すると信頼性が減少するという問題が生じる。一般にストレージの信頼性を増すにはRAIDを用いる。RAIDの中でもRAID6は2つのパリティで2台までの故障に耐える。しかし、大規模ストレージでは2耐故障のRAID6でも不十分である。現在、3台以上の故障に耐える技法にはRAID2と階層型RAIDのHiRAID(全RAIDによる多階層RAIDシステム)があるが、RAID2はシステムが複雑になるため実用化がされていない。HiRAIDでは2階層のRAID5であるRAID55が3耐故障であるが、HiRAIDには問題点として容量効率の低さがある。

本研究では同耐故障で階層型RAIDよりも容量効率に優れるRAIDである2進RAIDおよび2進RAIDを拡張したNaryRAIDの実装を目的とする。

第2章では信頼性を高めるストレージ技術と仮想的大規模ストレージを構築するVLSDについて、第3章ではNaryRAIDを実装する方法について提案し、第4章では今後の課題について述べる。

2 関連研究

2.1 RAID [1][2]

RAID(レイド)とはRedundant Array of Inexpensive(Independent) Disksの略で、データを複数のディスクに分散して記録して、高速化や信頼性の向上を得るための技術であり、ディスクへのデータ配置や、データの冗長化の方法により0から6までの7つのレベルで定義されて、信頼性を増す効果があるものには1耐故障性のあるRAID1,3,4,5および2耐故障性のあるRAID6がある。

2.2 階層型RAID

階層型RAIDはRAIDで構築したディスク(論理ドライブ)を1つのディスクとして見立てそのディスクでさらにRAIDを構築するシステムのことである。構築した階層型RAIDの耐故障性は各層のRAIDのデータを修復不可能な台数の積から1台引いたものとなるが、容量効率は各層の容量効率の積となる。

HiRAIDの代表例として16台のディスクからなるRAID55を図1に示す。

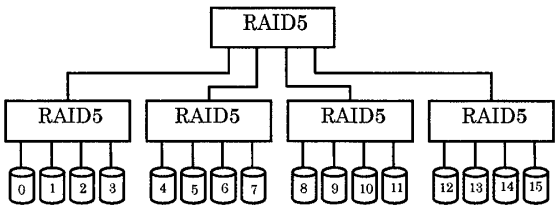

図1 RAID55

図1の16台からなるRAID55の容量効率は3/4*3/4=9/16=56%で、3耐故障である。

2.3 2進RAID[3]

同一ストライプ内で2進数の各桁に対応するグループを構成しRAID4形式でパリティをとる論理ディスクを構築する。2進RAIDでは2^n台のデータに対して2n台のパリティを用いるが、3耐故障であるためにはn>=3である必要がある。例としてn=3の2進RAIDの構成を図2に示す。8台のデータディスクd0～d7に対して6台のパリティディスクp0～p5がある。

d	0	1	2	3	4	5	6	7	0	1
2^0	0	1	0	1	0	1	0	1	p0	p1
2^1	0	0	1	1	0	0	1	1	p2	p3
2^2	0	0	0	0	1	1	1	1	p4	p5

図2　2進RAID(n=3)

ここで、p0,p1はディスク番号の0桁がそれぞれ0,1のディスクからなるグループ、p2,p3はディスク番号の1桁がそれぞれ0,1のディスクからなるグループ、p4,p5はディスク番号の2桁がそれぞれ0,1のディスクからなるグループである。

2.4 NaryRAID

2進RAIDでは2進数の各桁に対応するグループを構成していたものをN進数で構成できるよう

An Implementation of 3 Fault Tolerant RAID
Nakamura Yuji, Minoru Uehara
† Dept. of Information Sciences & Arts, Toyo Univ.

に拡張したもので、構成に必要なディスクの台数のパターンを多くすることができる。必要となるディスクの台数は N 進数では N^n 台のデータに対して Nn 台のパリティを用いる。3 耐故障であるためには 2 進 RAID と同様 n>=3 である必要がある。

2.5 VLSD [4]

VLSD(Virtual Large-Scale Disk)とは PC 教室などの教育環境の数百台の PC で安価で高信頼・大規模なファイルストレージを構築することを実現するためのツールキットである。

VLSD には仮想ディスクに関する vlsd、仮想ディスクの入出力に関する vlsd.io 、ディスクサーバに関する vlsd.server、仮想ディスクのテストを行うための test.vlsd、ディスクサーバのテストを行うための test.vlsd.server などがパッケージされている。

3 提案

本研究ではこの VLSD ツールを用いて NaryRAID を実装する。NaryRAID は RAID4 のパリティをとるグループで構成するため、VLSD の RAID4 クラスを N 進数の各桁に対応するグループで構成するように改良を行った。以下に今回の実装に関係するプログラムの説明を行う。

クラス DiskArray は複数ディスクからなるディスク・ラッパーの基底クラスである。DiskArray を作成し、標準では RAID1 を作成する。外部で引数 disks(要素ディスクの配列)を変更しても内部に影響しない。

クラス RAID は RAID の基底クラスで簡単な RAID1 の実装を DiskArray から引き継いでいる。引数の disks から初期故障のディスクを間引いて保持する。

クラス RAID4 は RAID のサブクラスで RAID4 の実装（ブロック単位ストライピングとパリティ専用ディスクの作成）を行う。

4 評価および検討

ここでは、JUnit のテストによって NaryRAID の 3 耐故障性能および既存の RAID との読み書き性能を比較、評価する。

まず NaryRAID にさまざまな組み合わせのディスクの故障を与えて 3 耐故障性を実現できているかを確認した。表 1 は NaryRAID に様々な故障を与えた結果で、故障ディスクに書かれた数字がデータディスクの故障の番号、p つきの数字がパリティディスクの故障の番号、「*」印は無故障を表している。いずれの場合も問題なく読み書きを行うことができ 3 耐故障性を実現しているといえる。

次に同台数の既存 RAID と NaryRAID との性能の比較を行った。表 2 は NaryRAID(N=2,n=3)とその必要台数である 93 台の RAID4,5,6 との読み書き性能の比較である。実験の結果、総ディスク台数が増えるとわずかながら他の RAID より書き込み時間が長くなったが、十分実用可能な範囲であると考えられる。

表 1 NaryRAID の故障テスト

進数[N]	次数[n]	台数	size	故障ディスク		
2	3	14	1MB	*	*	*
2	3	14	1MB	0	2	p1
2	3	14	1MB	6	p0	p3
2	3	14	1MB	p0	p2	p4
2	3	14	1MB	5	6	4
3	3	36	1MB	*	*	*
3	3	36	1MB	8	23	p7
3	3	36	1MB	0	8	23

表 2 NaryRAID と既存 RAID の読み書き時間

	N	n	Small Read[s]	Small Write[s]	Large Read[s]	Large Write[s]
NaryRAID	3	4	0.437	4.813	2.718	43.469
RAID4			0.437	2.438	2.734	38.172
RAID5			0.453	3.016	2.750	37.984
RAID6			0.416	4.375	2.687	40.610

5 現状

現段階で、VLSD での NaryRAID の実装は完了し、NaryRAID の性能評価として 3 耐故障性の確認と既存の RAID である RAID4、5、6 との読み書き性能の比較を行った。

しかし、階層型 RAID との読み書き性能における比較評価は階層型 RAID のテストがうまくいかなかったため行うことができなかった。

6 今後の課題とまとめ

今回実装した NaryRAID を用いることで容量効率のよい 3 耐故障を実現できる RAID を構築することができる。また NaryRAID は同じディスク台数の RAID4、RAID5、RAID6 とほぼ変わらない時間で読み書きが行えるので十分実用性のある RAID であると考える。

今後の課題としては更なる高速化や信頼性の向上などが考えられる。

参考文献

[1] ストレージ ETERNUS ストレージの基礎用語 : 富士通
http://storage-system.fujitsu.com/jp/term/beginner/
[2] RAID Level
http://homepage3.nifty.com/sekin/raid_level/raid_level.html
[3] 上原 稔 "2 進 RAID:もう一つの 3 耐故障 RAID", 信学技報 FIIS-08-228, (2008.3.7)
[4] チャイエリアント, 上原稔, 森秀樹 "PC 教室のための仮想的大規模ストレージの構築", マルチメディア、分散、協調とモーバイル(DICOMO 2007) シンポジウム論文集, pp.617-622, (2007.7.4-6)