Vol.2009-DPS-140 No.14
2009/9/11

NaryRAID の実装

中村祐司[†]　　上原稔[†]

今日，ブログなどの CGM の普及や，写真ないし動画投稿サイトの流行により，ネットワーク上の情報が増大している．これらの情報は多量のディスクを用いた大規模ストレージの運用によりデータを保存しているが，ストレージのディスクが増えると信頼性が低下するという問題が発生する．ストレージの信頼性を向上させる技術として RAID があるが既存の RAID では 1 または 2 台の故障までしか修復できず信頼性が不十分で，階層型 RAID を用いれば十分な信頼性を得られるが容量効率が低くなるという問題点がある．本研究では非階層で 3 耐故障を実現し階層型 RAID より容量効率に優れる RAID として提案された 2 進 RAID と 2 進 RAID を拡張した NaryRAID を実装及び評価について述べる．

An Implementation of NaryRAID

YUJI NAKAMURA[†]　MINORU UEHARA[†]

Today, CGM(Consumer Generated Media) such as blog is spread. In addition, photo and video sharing sites become popular. In this way, the amount of information in network increases drastically. Such information is usually stored into a large-scale storage which consists of many disks. In such a storage system, the larger the number of disks is, the lesser the reliability of the system. RAID is usually used to improve the reliability of a storage. However, it is not enough because conventional RAID repair only either one or two faults. Therefore, we require more than 2 fault tolerant RAID. If hierarchical RAID is employed, more than 2 fault tolerance can be realized but the capacity efficiency is decreased. We have proposed NaryRAID which realizes 3-fault tolerant without hierarchical organization. The capacity efficiency of NaryRAID is larger than that of hierarchical RAID. In this paper, we describe an implementation of NaryRAID and evaluate its performance.

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．その結果，Amazon S3, Nirvanix などのオンラインストレージサービスが誕生した．SOA の視点から見てもオンラインストレージサービスはその他の Web サービスの基盤となる．本論文では，S3 のようなオンラインストレージサービスを大規模ストレージサービスと名づける．

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．商業製品では RAID6 が主流になりつつある．しかし，大規模ストレージサービスでは 2 耐故障の RAID6 でも不十分である．Accusys はパリティを 3 つに増やし 3 耐故障とした RAID6TP を開発したが，普及はしていない．

従来, 3 台以上の故障に耐える技法には RAID2 と階層型 RAID の HiRAID[1](全 RAID による多階層 RAID システム)しかなかった. 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 の実装を及び評価について述べる．

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述べる．第 4 章では実装した NaryRAID の評価を述べる．最後にまとめを述べる．

[†] 所属
東洋大学情報工学科
Dept. of Information and Computer Sciences, Toyo University, Japan

ⓒ2009 Information Processing Society of Japan

2. 関連研究

2.1 階層型 RAID

階層型 RAID とは 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 RAID を構築することである．階層型 RAID の中でも RAID1 と RAID0 の組み合わせた構成が一般的にも用いられていて特別に RAID 1+0(RAID 10)/RAID 0+1(RAID 01)と呼ばれている．

RAID10 は RAID1 で構築した論理ドライブで RAID0 を構成する．この組み合わせを 1 つのアレイに構成することで，RAID1 によるデータ二重化と，RAID0 の高速化を合わせて実現できる．しかし RAID0 には信頼性の向上の効果がないため RAID0 と組み合わせた階層型 RAID では高速化の効果が得られることはできても耐故障性を向上させることはできない．3 台以上の故障に耐えるには HiRAID による階層型 RAID で信頼性を向上さる RAID クラスでの組み合わせが必要となる．HiRAID で構築した階層型 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 台引いたものとなり，容量効率は各層の容量効率の積となる．

2.2 NaryRAID

NaryRAID[1][3]は我々が開発した 3 耐故障 RAID 方式である．既存 RAID クラスでは，エラー訂正符号に基づく RAID2 に 3 以上の耐故障性がある．しかし，エラー訂正符号の計算は複雑であること，およびビット単位の計算であるため性能を重視する RAID には適さない．おそらく後者が最大の原因と考える．以上のことから 3 耐故障 RAID もブロック単位のパリティ方式でなければならない．

RAID6 は RAID5 より複雑であるが，RAID2 よりは単純である．RAID6 には P+Q 方式と 2D-XOR 方式がある．2D-XOR 方式は RAID5 と同様にストライプ内で水平パリティ(PH)をとるほか，別のストライプに対して対角パリティ(PD)をとる．PD がボトルネックとなり限界性能が低い．P+Q 方式ではストライプ内で 2 種のパリティをとるためボトルネックがない．そのためほとんどの商用 RAID6 は P+Q 方式を採用している．

しかし，P+Q 方式を 3 耐故障に拡張することは容易ではない．そこで，本論文では，ボトルネックが発生しないように同一ストライプ内で複数のグループを構成しパリティをとる．グループは，N 進数の各桁に対応する．よって，この手法を N 進数(N-ary number)に基づく RAID として NaryRAID と名付ける．NaryRAID は基数 N とその次数 n で特徴付けられる．そこで，NaryRAID(N,n)と表記する．例えば N=2,n=3 の場合の NaryRAID は NaryRAID(2,3)とあらわす．

ここで，もっとも単純な NaryRAID である 2 進 RAID を例に説明する．図 1 に NaryRAID(2,2)の場合の 2 進 RAID を示す．4 台のデータディスク d0～d3 に対して 4 台のパリティディスク p0～p3 がある．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（これをパリティグループと名づける），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である．このように NaryRAID は RAID4 に基づく．

disk	d0	d1	d2	d3	0	1
2^0	0	1	0	1	p0	p1
2^1	0	0	1	1	p2	p3

図 1　2 進 RAID(n=2)
Figure 1 BinaryRAID(n=2)

n=2 の場合，3 台故障時にデータディスクとそのすべてのパリティディスクが同時に故障する可能性があるため 3 耐故障でない．ゆえに，すべてのデータディスクが必ず 3 つ以上のパリティグループに属せばよい．n>=3 のとき，その条件を満たす．図 2 に NaryRAID(2,3)のときの 2 進 RAID の構成を示す．

disk	d0	d1	d2	d3	d4	d5	d6	d7	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)
Figure 2 BinaryRAID(n=3)

パリティはグループ内で 1 つの故障まで修復できる．ディスクが必ず異なるグループに属せば，パリティによっていずれの故障も修復できる．本方式では N 進数に基づきグループを組織する．N 進 RAID では N^n 台のデータに対して Nn 台のパリティを用いる．よって，その容量効率は以下の式で表せる．

$$\frac{N^n}{N^n + Nn} \tag{1}$$

3 耐故障である NaryRAID と RAID55 を容量効率の面で比較する．NaryRAID の容量効率は式(1)の通りである．対して RAID55 の容量効率は上位層と下位層の数が等しいとき最大となる．すなわち全体の台数を D とするとそれぞれ$\sqrt{D}$となり容量効率は $D - 2\sqrt{D} + 1 / D$ で求めることができる．表 1 に NaryRAID(N=2)である 2 進 RAID と RAID55($D=2^n+2n$)の場合を示す．表からわずかながら NaryRAID の方が RAID55 より容量効率が優れていることが分かる．

ⓒ2009 Information Processing Society of Japan

表 1　2 進 RAID と RAID55 の容量効率
Table 1 The capacity efficiency of BinaryRAID and RAID55

n	総台数(D)	2 進 RAID	RAID55
3	14	57.14%	53.69%
4	24	66.67%	63.34%
5	42	76.19%	71.52%
6	76	84.21%	78.37%
7	142	90.14%	83.92%
8	272	94.12%	88.24%
9	530	96.60%	91.50%

2.3 VLSD

VLSD(Virtual Large-Scale Disk)[5]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 はすべてのプログラムが Java 言語で書かれており，Java(J2SE)が動作するプラットフォームなら動作させることができる．また Java の long 仕様により最大 64 ビットのディスク容量の大規模ストレージをクライアント PC の空き容量を束ねて構築できる．

VLSD ツールキットには以下のクラスが含まれる．

FileDisk

単一ファイルによる固定容量ディスク．論理的な容量と物理的な容量は正確に一致する．Java.io.RandomAccessFile により実装される．

VariableDisk

単一ディスクにより容量可変ディスクを作成するラッパー．8KB を単位とする多分木で管理する．葉ノードには 8KB のデータが格納される．中間ノードには 1024 個の 64ｂ(8B)ポインタが格納される．ノードには必要に応じて割り当てられる．6 階層で 8EB-1 まで拡張できる．データ以外の管理情報が保存されるため物理的な容量は 0.1%増加する．容量可変ディスクを実現するため，Disk インターフェースには容量を追加する API が定義されている．

NBDDisk/NBDServer

NBD デバイスのクライアント．NBDServer と NBD プロトコルで通信する．その他の NBD サーバ実装とも通信できる．

RmoteDisk/RemoteServer

遠隔デバイスのクライアント．RMI プロトコルで通信する．RemnoteDisk に対応するサーバは DiskServer である．

SecureRemoteDisk/SecureRemoteServer

アクセスキーによる安全な遠隔デバイスのクライアント．RMI プロトコルで通信する．SecureRemoteDisk に対応するサーバは SecureDiskServer である．

JBOD

複数のディスクを直列に連結したディスク．冗長性がなく．容量増のために用いられる．各ディスクの容量は一様でなくてもよい．ストライピングを行わないため容量は単純に総和となる．例えば, 100GB, 120GB, 160GB を連結すると 100+120+160=380GB になる．JBOD に対して連続的に逐次アクセスをすると特定の部分ディスクに負荷が集中する．

RAID*n*(*n*=0,1,4,5,6)

各 RAID クラスの実装．RAID0 は HW RAID と異なり，JBOD と有意な差はない．RAID4,5 は 1 耐故障である．RAID5 は HW RAID と異なり，RAID4 との有意な差はない．RAID6 は 2 耐故障である．P+Q 方式を採用している．

VotedRAID1

RAID1 に似ているが，多数決で任意故障をマスクする．書込み操作はすべてのディスクに複製される．読み取り操作はすべてのディスクに複製され，その結果を多数決する．多数決のため最低 3 台のディスクを必要とする．稼動ディスクが 2 台以下になると正しく多数決できなくなる．

これらのクラスは自由に組み合わせることができる．例えば，RAID6 を 2 段階で組み合わせると RAID66 を構築できる．

3. NaryRAID の実装

VLSD に NaryRAID のクラスを追加し実装を行った．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し実装を行う．

以下に今回の実装に関係するプログラムの説明を行う．

RAID4

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．用意された要素ディスクで一番後ろのディスク番号のものをパリティディスクとして構築する．

NaryRAID

クラス NaryRAID は RAID のサブクラスで NaryRAID の実装である．用意された要素ディスクから指定した基数とレベルからデータディスクの台数とパリティディスクの台数を求め，ディスク番号 0 からデータディスクとして，データディスクの最後の

ⓒ2009 Information Processing Society of Japan

番号に 1 足したディスク番号からパリティディスクとして NaryRAID を構築する．NaryRAID には次の以下のメソッドを持つ．

repair

repair メソッドは故障した NaryRAID の要素ディスクの番号を格納した配列から修復可能な手順を導きだし，読み込みデータを計算する．

make_depends

make_depends メソッドはパリティディスクの番号からそのパリティディスクを構成するデータディスクの集合を求める 2 重配列 dep[j][]を作成する. dep[j][]はパリティディスク[j]を構成するデータで中にはデータディスク番号が格納される．

make_repairs

make_repairs メソッドはデータディスクの番号からそのデータディスクが持つパリティを求める 2 重配列 rep[no][k]を作成する．rep[no][k]はデータディスク[no]が作るパリティで中にはパリティ番号が収納される．

4. 評価

はじめに NaryRAID が n=3 において常に 3 耐故障であることを実装したプログラムを用いて証明する. 総ディスク台数 14 台 1 つのディスクサイズ 1MB の NaryRAID(2,3)と総ディスク台数 36 台の NaryRAID(3,3)の時に 3 台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 2 は任意に 3 台の故障を与えたときの結果である．故障させたディスクは故障ディスクの項目に示した数字で，頭に d がついたものがデータディスク，頭に p がついたものがパリティディスクのディスク番号の故障を，*印は故障していないことを表している．実験の結果いかなる場合も問題なく読み書きを行えることができ 3 耐故障性能を実現していることが確認できた．

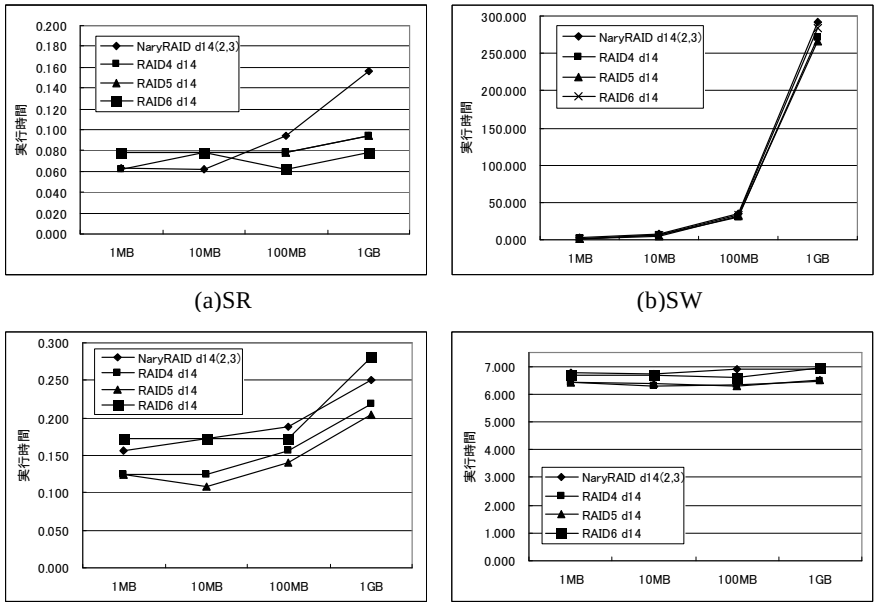
表 2 NaryRAID の故障テスト
Table 2 Test of failed NaryRAID

進数[N]	次数[n]	台数	size	故障ディスク		
2	3	14	1MB	*	*	*
2	3	14	1MB	d0	d2	p1
2	3	14	1MB	d6	p0	p3
2	3	14	1MB	p0	p2	p4
2	3	14	1MB	d5	d6	d4
3	3	36	1MB	*	*	*
3	3	36	1MB	d8	23	p7
3	3	36	1MB	d0	d8	d23

次に無故障時の NaryRAID の基本性能を評価する．なお，本評価ではディスク台数を一定とし，そのサイズを変更して性能との関係を調査する．

同じディスク台数の NaryRAID と既存 RAID との無故障時の読み書き時間の評価を行った結果を図 3 に示す. 実験では 1 つのディスクのサイズをそれぞれ 1MB, 10MB, 100MB，1GB でディスク台数を NaryRAID(2,3)の必要台数である 14 台に統一し NaryRAID と既存 RAID クラスである RAID4，RAID5，RAID6 との読み書き性能の比較を行った． SmallRead(SR)/Write(SW)は同一ディスクに対する読み書きであり，LargeRead(LR)/Write(LW)はグループ全体におよぶ読み書きである．実験の結果 NaryRAID の SW が他の RAID クラスと比べてわずかに遅れるがほとんど同じ時間で読み書きが行えることがわかる．

またディスクサイズの増加における読み書き時間の変化は，いずれの RAID クラスも SW のみディスクサイズに比例して実行時間が増加して SR，LW，LR はサイズにおける所要時間の変化がないことが分かる．



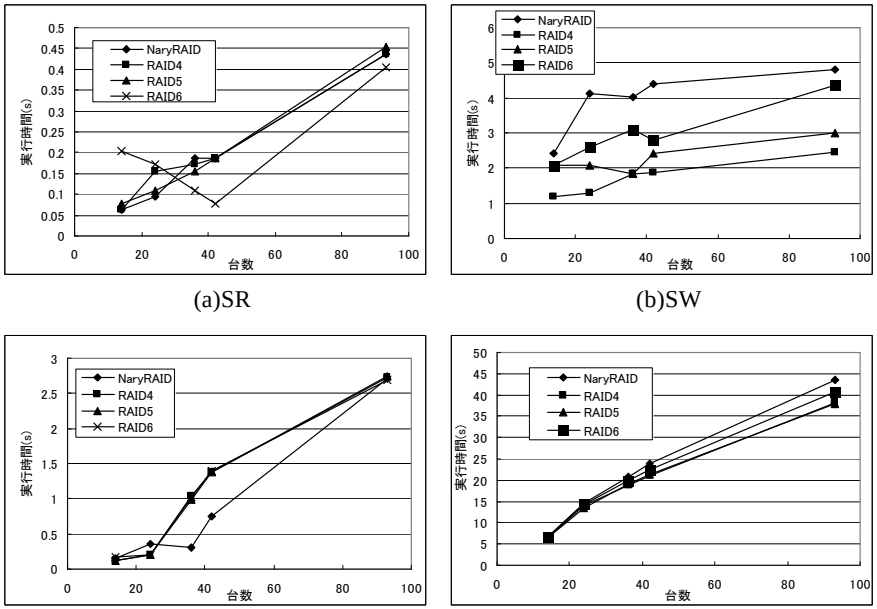
(a)SR (b)SW (c)LR (d)LW

図 3 ディスクサイズ増加時の性能評価
Figure 3. The performance evaluation in case of increasing the size of a disk

ⓒ2009 Information Processing Society of Japan

次に無故障時の NaryRAID の基本性能を，サイズを一定，ディスク台数を可変として調査した．

総ディスク数の増加における性能の評価を行った．図 4 は NaryRAID(N,n)における(N,n)の組がそれぞれ(2,3)，(2,4)，(3,3)，(2,5)，(3,4)，の必要台数である 14 台，24 台，36 台，42 台，93 台における NaryRAID と RAID4，RAID5，RAID6 の無故障時の読み書き時間の結果である．いずれも 1 つのディスクサイズは 1MB に固定して実験を行った．総ディスク台数が増えていくと SW 及び LW において NaryRAID は他の RAID より遅くなっていくのが読み取れる．LW の変化は台数が増えるごとに RAID4,5 よりもわずかに RAID6 が遅くなっていき，さらにそれよりも NaryRAID が遅くなっていくことが分かる．



(a)SR (b)SW

(c)LR (d)LW

図 4. ディスク台数増加時の性能評価
Figure 4. The performance evaluations in case of increasing #disks

次に 1 耐故障である RAID4,5 と 2 対故障である RAID6 を可能な台数故障させた場合と，同じ台数故障させた NaryRAID との性能を比較した．表 3 は無故障時，表 4 は 1 台故障時，表 5 は 2 台故障時の NaryRAID(3,4)の時と，その時と同じディスク台数の時の読み書き時間の比較である．全体的に故障が増えるごとにわずかな速度の低下が見られた．

表 3. 無故障時の NaryRAID と既存 RAID の性能比較
Table 3. The performance comparison of NaryRAID with conventional RAID in case of no failure

	台数	size	SR[s]	SW[s]	LR[s]	LW[s]
NaryRAID	93	1MB	0.375	5.203	2.625	42.844
RAID4	93	1MB	0.359	2.078	2.688	37.562
RAID5	93	1MB	0.375	3.032	2.656	37.672
RAID6	93	1MB	0.375	3.875	2.656	40.374

表 4 1 台故障時の NaryRAID と既存 RAID の性能比較
Table 4. The performance comparison of NaryRAID with conventional RAID in case of single failure

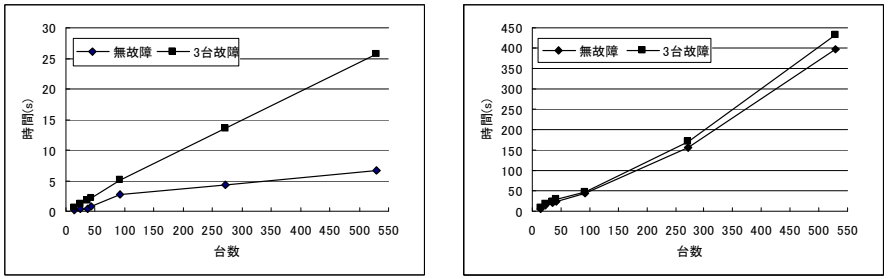
	台数	size	故障	SR[s]	SW[s]	LR[s]	LW[s]
NaryRAID	93	1MB	6	0.390	6.000	3.094	44.625
RAID4	93	1MB	6	1.188	3.500	3.218	38.875
RAID5	93	1MB	6	0.406	2.609	3.219	39.063
RAID6	93	1MB	6	0.438	4.953	4.312	43.875

表 5 2 台故障時の NaryRAID と既存 RAID の性能比較
Table 5. The performance comparison of NaryRAID with conventional RAID in case of 2 failures

	台数	size	故障		SR[s]	SW[s]	LR[s]	LW[s]
NaryRAID	93	1MB	80	45	0.406	3.938	3.953	46.250
RAID6	93	1MB	80	45	0.422	5.781	8.688	52.922

次に，ディスク台数の増加によるディスク無故障時と 3 台故障時の NaryRAID 読み書き時間の比較を行った．図 5 より LR，LW においてディスク台数が増加するほど故障時の読み書き速度が低下することが分かる．

ⓒ2009 Information Processing Society of Japan



(a)LR (b)LW

図 5. 台数増加における NaryRAID の故障時と無故障時の性能比較
Figure 5. The performance of NaryRAID in case of 3 failures

最後に, 総ディスク台数が同じで進数, 次数の異なる NaryRAID の性能を比較する. NaryRAID の総台数は N^n+Nn で求められるので，2 進 8 次(2,8)の時と 4 進 4 次(4,4)の時に同じ台数で異なる組み合わせ条件を満たす. 表 6 は 2 進 8 次(2,8)の時と 4 進 4 次(4,4)の時の読み書き時間の結果である．全体的に 2 進 8 次の時よりも 4 進 4 次のほうが読み書きの時間が早く故障時の速度低下も減少した.

表 6 NaryRAID(2,8)と NaryRAID(4,4)の性能比較
Table 6. The performance comparison of NaryRAID(2,8) with NaryRAID(4,4)

進数 [N]	次数 [n]	台数	size	故障ディスク			SR[s]	SW[s]	LR[s]	LW[s]
2	8	272	1MB	*	*	*	1.219	8.375	4.312	156.188
2	8	272	1MB	43	138	172	0.844	9.094	13.546	171.907
4	4	272	1MB	*	*	*	0.703	5.875	4.250	121.953
4	4	272	1MB	109	144	167	0.875	8.375	9.734	131.157

5. まとめ

本論文では NaryRAID の実装し評価を行った．評価では NaryRAID の 3 耐故障テストと RAID4,5,6 との比較を行い．NaryRAID はいかなる場合の 3 台の故障でも問題なく読み書きを行うことができ 3 耐故障を満たしていることが確認できた．無故障時の読み書きの時間は同じ台数の既存 RAID と NaryRAID ともにほとんど変わりがなかった．1 つのディスクのサイズを大きくすると既存 RAID と NaryRAID とも SW がサイズに比例して増える傾向にありわずかながら既存 RAID より速度が低下している．要素ディスク数が増えると既存 RAID，NaryRAID ともに LW が増える傾向にあり，RAID4,5 より RAID6 が，RAID6 より NaryRAID が，わずかに速度が低下した．故障時と無故障時の読み書きの時間の比較では，総ディスク台数が増えると，無故障時と比べ故障時の LR と LW に低下が見られた．全体的に規模が大きくなると既存 RAID よりわずかな速度の低下がみられたが，十分実用可能な範囲であると考えられる．

今後の課題として，今回の実験では階層型 RAID の評価がうまく行えなかったが，3 耐故障 RAID の性能評価として，同耐故障である階層型 RAID の RAID55 との性能比較を行う必要がある．また，文献[4]では複数パリティを持った NaryRAID について考察しているが，その実装を行う．

謝辞

本研究は科研費基盤(C)「PC グリッドによる高信頼・高効率な分散仮想ストレージの研究(19500066)」により援助されています.

参考文献

[1] Sung Hoon Baek, Bong Wan Kim, Eui Joung Joung and Chong Won Park: “Reliability and Performance of Hierarchical RAID with Multiple Controllers,” In Proc. of 20th annual ACM symposium on principles of distributed computing, pp.246-254, (2001)
[2] 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)
[3] Katsuyoshi Matsumoto, Minoru Uehara: "N-nary RAID: 3-resilient RAID based on an N-nary number", In Proceedings of 23rd International Conference on Advanced Information Networking and Applications(AINA2009), pp.249-255, (2008.5.26)
[4] Minoru Uehara: "Combining N-ary RAID to RAID MP", ITIS2009, (TBA)
[5] 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)

ⓒ2009 Information Processing Society of Japan