M-015

NaryRAID の評価

The evaluations of NaryRAID

中村 祐司† 上原 稔†
Yuji Nakamura Minoru Uehara

1. はじめに

低コストの大容量ストレージに対する要求は非常に高い．我々は，空容量を集約してこのようなストレージを構築するために，ディスクレベル分散型ストレージを構築するためのツールキット VLSD (Virtual Large Scale Disks) を開発した．大規模ストレージでは，耐故障性が重要である．よって，ストレージの信頼性を高める技術が重要である．

ストレージの信頼性を高める技術の一つに RAID[1][2] がある．RAID には，0 から 6 まで 7 つの基本クラスがある．しかし，基本クラスのみでは RAID6 の 2 耐故障までしか実現できない．大規模ストレージを構成するには少なくとも 3 以上の耐故障性を実現する必要がある．

3 耐故障 RAID を構成する最も容易な方法は階層 RAID (HRAID)である．図 1 に 3 耐故障である RAID55 を示す．

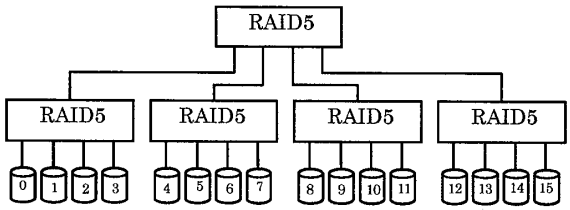


図 1 RAID55

RAID55 は 2 階層の RAID5 である．階層 RAID は，構成が容易である利点があるが，容量効率が良くない．そこで，我々は独自に N 進数に基づくパリティグループを構成することで 3 耐故障を実現する NaryRAID を提案し，VLSD で実現した．NaryRAID は同じ 3 耐故障である RAID55 より容量効率に優れるが，2 耐故障である RAID6 には大きく劣っている．本論文では，容量効率や信頼性，性能などの観点から RAID6 や RAID55 に変わり NaryRAID が有用であるストレージシステムの規模について考察する．

本文の構成は以下の通りである．2 節では NaryRAID について述べる．3 節では VLSD について述べる．4 節では評価を行い，最後に結論を述べる．

2. NaryRAID

NaryRAID[3][4]は我々が開発した 3 耐故障 RAID 方式で，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．

RAID4 を用いて 2 進 RAID の基本アイデアを説明する．2 進 RAID は 2 進数に基づく．2 進数の桁数を n とすると，n=3 における 2 進 RAID の構成を図 2 に示す．例えば，パリティ p0 はディスク d*i* のディスク番号 *i* の 1 桁目が 0 であるディスクグループのパリティである．

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 BinaryRAID(n=3)

n<3 のとき 2 進 RAID は 3 耐故障ではない．その原因は，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である．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．

N 進 RAID は 2 進 RAID を単純に N 進数に拡張したものである．N 進 RAID では基数 N と最大データディスク数 Nn を示すレベル n で特徴づけることができる．よって，N 進 RAID のクラスを NaryRAID(N,n)と示すこととする．2 進 RAID は NaryRAID(2,n)である．図 3 に NaryRAID(3,3)の構成を紙面の都合上省略して示す．

	0	1	2	3	4	5	6	…	25	26	0	1	2
0	0	1	2	0	1	2	0	…	1	2	0	1	2
1	0	0	0	1	1	1	2	…	2	2	3	4	5
2	0	0	0	0	0	0	0	…	2	2	6	7	8

図 3 NaryRAID(3,3)

NaryRAID(3,3)のデータディスク数は 27，パリティディスク数は 6 である．一般に NaryRAID(N,n)のデータディスク数は Nn，パリティディスク数は Nn である．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．容量効率を比較する．結果を図 6 に示す．ここで，x 軸は n を示し，RAID55(2)，RAID55(3)はそれぞれ

† 東洋大学 Toyo University

N=2，3 としたときの NaryRAID と等しいディスク数の RAID55 における容量効率である．なお，RAID55 の容量効率は構成により異なるため，容量効率を最大化する構成で比較した．

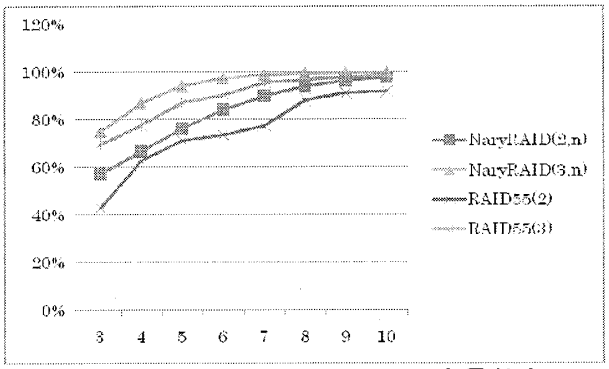


図 4　NaryRAID と RAID55 の容量効率

図から明らかなように n が増えるほど容量効率は高まる．また，常に NaryRAID が RAID55 を上回る．よって，本方式は多数のディスクを用いる大容量ストレージに適する．

3. VLSD

VLSD (Virtual Large Scale Disks)[5]は Java で記述された大規模ストレージ構築用ツールキットである．この中には様々なストレージクラスが実装されており，これらを組み合わせることで有用なソリューションを提供できる．

VLSD の主なクラスは以下のとおりである．

RAID0,3,4,5,6

各 RAID クラス．

SingleRAID

単一ディスクを RAID にみせるラッパー．

StripeDisk

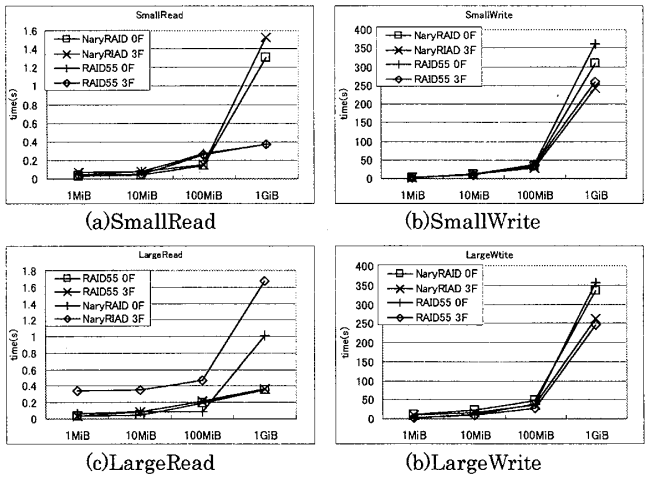
RAID のストライプに合わせて，データをアクセスする．これにより RAID 中の特定のデータディスクのみをアクセスできる．分散パリティでない RAID0,1,3,4 と併用する．StripeDisk は RAID 中のデータディスクに対してのみ使用できる．パリティディスクを読み取ることは原理的に可能だ（すべてのデータディスクの排他的論理和として表現できる）が，データディスクを変更せずにパリティディスクを修正することはできないため，パリティディスクへの書き込みを表現することはできない．

NaryRAID

オールインワンで実装された NaryRAID．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．

4. 評価

ここでは，JUnit のテストによって 3 耐故障 RAID である NaryRAID と RAID55 との読み書き性能を比較，評価する．図 5 は NaryRAID(N=2,n=3)と 4 台の RAID5 の 2 階層からなる 16 台の RAID55 との読み書き性能の比較である．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．どちらも無故障時と 3 台故障時の評価を行った．実験の結果，リードにおいてはディスクサイズが大きくなると RAID55 の方が早くなるが，ライトにおいては NaryRAID の方が優れている．故障時には NaryRAID, RAID55 ともに無故障時よりも書き込みが早くなる．



(a)SmallRead　(b)SmallWrite

(c)LargeRead　(b)LargeWrite

図 5　NaryRAID と既存 RAID の読み書き時間

5. まとめ

本論文では 3 耐故障 RAID である RAID55 と NaryRAID の比較，評価を行った．容量効率においては NaryRAID は常に RAID55 を上回る．また読み書きの時間においては NaryRAID は RAID55 と比べ，読み込みが遅く書き込みが早い．読み書き時間はディスクサイズを大きくしていくとライトにかかる時間が大きく占めるようになる．よって，ディスクサイズが大きいディスクを多量に用いる大規模ストレージにおいて NaryRAID は RAID55 より有用であるといえる．

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]P.M. Chen, E.K. Lee, G.A. Gibson, R.H. Katz, and D.A. Patterson: “RAID: High-Performance, Reliable Secondary Storage,” ACM Computing Surveys, Vol. 26, No. 2, pp.145-185, June 1994

[3]K. Matsumoto, M. 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: "3 Faults Tolerant Orthogonal RAID for Large Storage", NBiS2010, TBA

[5]Minoru Uehara: "A Toolkit for Virtual Large-Scale Storage in a Learning Environment", In Proc. of 21th International Conference on Advanced Information Networking and Applications Workshops/Symposia 2007, Vol. 1, pp.888-893, (2007.5.23)