Vol.2011-DPS-149 No.4
2011/11/24

大規模仮想ディスクにおける複数パリティ直交 RAID

上原 稔[†]

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。しかし、大容量ストレージの信頼性を高めるには 3 以上の耐故障性を持つ RAID が必要になる。過去の研究において、我々は、直交 RAID に基づく 3 耐故障 RAID として MeshRAID を提案し、それを大規模仮想ディスク(VLSD)によって実装した。MeshRAID はディスク故障だけでなく配線故障にも耐える。また、VLSD のクラスライブラリを組み合わせる簡易実装である複合 RAID(Compisite RAID)を提案し、それによる MeshRAID の実装も行った。本論文では、MeshRAID を構成する RAID に複数のパリティを持ち、複数の故障に耐える RAID クラスを採用した MeshRAID MP を提案する。具体的には、MeshRAID MP としては最も基本的な MeshRAID2P について検討する。本論文で採用する 2 耐故障 RAID は水平パリティと対角パリティの 2 つを持つ RAID DP である。

Orthogonal RAID with Multiple Parties in Virtual Large-Scale Disks

Minoru Uehara[†]

Recently, the demand of low cost large scale storages increases. We developed VLSD (Virtual Large Scale Disks) toolkit for constructing virtual disk based distributed storages, which aggregate free spaces of individual disks. However, in order to construct large-scale storage, more than or equal to 3 fault tolerant RAID is important. In the previous work we proposed MeshRAID that is 3 fault tolerant orthogonal RAID. And, we implemented MeshRAID using VLSD. MeshRAID is not tolerant of only disk fault but also tolerant of string fault. In addition, we also proposed Composite RAID, which is an easy implementation of complex RAID system by combining VLSD classes. We also implemented Composite RAID based MeshRAID. In this paper, we propose MeshRAID MP(Multiple Parities), which has multiple fault tolerance using multiple parities. Specifically, we implement MeshRAID2P. Furthermore, we employ RAID DP, which has row parity and diagonal parity, as 2FT RAID.

[†]東洋大学 総合情報学部
Faculty of Information Sciences and Arts, Toyo University.

1. はじめに

ストレージ技術の進歩とクラウドの普及により、オンラインストレージに対する要求はますます高まっている。クラウドのような大規模ストレージでは、大量のディスクを使用する。しかし、ディスク単体の信頼性がどれほど高まろうと、その数が増えるほど反比例して信頼性は低下する。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 がある。普及している RAID[1)2)]では、2 耐故障の RAID6 が最高である。しかし、大量のディスクを用いる大規模ストレージでは 2 耐故障でも十分とは言えない。3 耐故障以上の信頼性が必要となる。

基本的に 3 以上の耐故障 RAID は RAID の基本クラスでは実現できない。3 耐故障 RAID を実現するには階層 RAID[3)]がある。階層 RAID では 2 層の RAID5 が 3 耐故障となる。また、我々は階層 RAID より容量効率の優れた NaryRAID[5)6)7)]を提案した。また、ディスク故障だけでなく配線故障にも耐える MeshRAID[11)14)15)]も提案した。中でも MeshRAID は容量効率こそ階層 RAID と等しいものの、配線故障にも耐え、最も高い信頼性を持つ。

MeshRAIDには様々なバリエーションがある。基本的なMeshRAIDは2階層のRAID4で構成される MeshRAID44 である。MeshRAID は分散パリティの RAID5 でも構成できる。MeshRAID55 は 2 階層の RAID5 で構成される。MeshRAID44/55 はいずれも 3 耐故障である。

3 耐故障 RAID で十分な規模は比較的小さい。中規模なストレージに適している。ストレージの規模がもっと大きくなると 3 耐故障でも十分ではない。

MeshRAID は階層を増やすこともできる。MeshRAID444 は 3 階層の RAID4 で構成される。MeshRAID555 は 3 階層の RAID5 で構成される。RAID4/5 に基づく MeshRAID の耐故障性は階層数 n に依存する。n 階層の MeshRAID の耐故障性は 2^n-1 である。しかし、各層のディスク数を N とすると容量効率は $((N-1)/N)^n$ となり、n に反比例する。

MeshRAID の耐故障性を増やすにはパリティ数を増やす方法もある。RAID4/5 は 1 パリティで 1 耐故障を実現するが、RAID6 は 2 パリティで 2 耐故障を実現できる。RAID6 と同様に対角 RAID は、水平パリティと対角パリティの 2 つのパリティで 2 耐故障を実現する。このような 2 パリティ RAID を用いると耐故障性は 3^n-1 となる。一般的には m パリティ m 耐故障 RAID を用いると MeshRAID の耐故障性は $(m+1)^n$-1 となる。それでも容量効率は $((N-m)/N)^n$ となる。

ここで、両指標を同時に考慮するために容量当たりの耐故障性 E を考える。E は耐故障性と容量効率の積となる。ここで、ディスク総数を D とすると $N=D^{1/n}$ である。容量効率は $(1-m/N)^n \fallingdotseq 1-nm/N^n$ で近似できる。よって、$E \fallingdotseq ((m+1)^n-1)(1-nm/D)$ となる。この式から、m より n の方が相対的に支配的であるといえる。表 1 に D=768 の場合にお

ける m, n と E の具体的な関係を示す。この表から E に関して m, n はトレードオフの関係にあると言える。

表 1. E と m, n の関係(D=768)
Table 1. The relationship of E to m and n(D=768)

n \ m	1	2	3	4
1	1.0	2.0	3.0	4.0
2	2.8	6.9	11.9	17.4
3	4.9	12.2	18.7	21.3
4	6.1	10.4	6.5	1.0

本論文では、パリティ m に注目する。複数のパリティを持つ要素 RAID で MeshRAID を構成する。具体的には、2 パリティの対角 RAID を用いて MeshRAID を実装し、VLSD(Virtual Large-Scale Disks)[4)9)]を用いて評価を行う。まだ、3 パリティ 3 耐故障 RAID は一般的でないため、現実的に構成可能な MeshRAID MP は MeshRAID 2P に限られる。

本文の構成は以下の通りである。2 節で関連研究として既存の 3 耐故障 RAID について述べる。3 節では VLSD について述べる。4 節では、MeshRAID MP の概念と VLSD による実装法について述べる。5 節では、その評価を行う。最後に結論を述べる。

2. 関連研究

RAID の規模を増加させる技法に階層型 RAID がある。階層型 RAID では、通常の RAID より信頼性を向上させることが可能となる。例えば、通常の RAID5 は 1 耐故障だが、RAD55（2 階層の RAID5）は 3 耐故障である。RAID55 の信頼性は RAID6 を上回る。一般に n 階層 RAID5 の耐故障性は 2^n-1 である。しかし、RAID55 の容量効率は RAID6 より低い。

階層型 RAID では、階層を増やすほど信頼性が高まり、同時に規模も拡大する。例えば、n 層の RAID があり、各層が k 個の RAID5 からなるとすると、その規模は k^n であり、信頼性は 2^n-1 である。しかし、階層が増えると、容量効率は低下し、経済性が急速に劣化する。RAID の経済性は容量効率で決定される。k が十分大きいとき、その容量効率は以下の式で表される。

$$\left(\frac{k-1}{k}\right)^n = \left(1-\frac{1}{k}\right)^n \approx 1-\frac{n}{k}$$

ゆえに、k>>n であることが望ましい。本研究では、複製方式との分岐点である容量効率 50%以上を目標とする。しかし、普及品 RAID では、k が小さく規模の拡大が困難である。

階層 RAID の要素 RAID のパリティ数 m が 1 より大きい場合、耐故障数も大きくなる。一般に m パリティ RAID の耐故障数は m 以下である。2 パリティで 2 耐故障の RAID の例は RAID6, RAID DP である。これらは RAID5 に代わり市場に普及しつつある。今後は m パリティ m 耐故障 RAID が普及すると予想される。このような m パリティ RAID からなる階層 RAID の耐故障性は $(m+1)^n$-1 である。また、その容量効率は以下の式で表される。

$$\left(\frac{k-m}{k}\right)^n = \left(1-\frac{m}{k}\right)^n \approx 1-\frac{nm}{k}$$

容量効率 50%以上であるためには k>2nm でなければならない。

我々は RAID55 より容量効率の優れた 3 耐故障 RAID 方式として、NaryRAID を提案した[5),6)]。NaryRAID では、ディスク番号の N 進数における各桁の数値が等しいディスクを集めてパリティグループを構成する。各ディスクが少なくとも 3 つのグループに所属すれば 3 耐故障となる。NaryRAID の基数を N とし、データディスク数 D を N^n としたとき、パリティディスク数 P は Nn である。NaryRAID は基数 N と次数 n によって特徴づけられる。よって、一般化して NaryRAID(N,n)と表す。図 1 に基数 2、次数 3 の NaryRAID(2,3)を示す。NaryRAID(2,3)では、8 台のデータディスクと 6 台のパリティディスクで構成される。基数 N が小さいと容量効率はあまり大きくないが、基数 N が十分大きければ P は無視できるほど小さくなる。よって、容量効率は高い。しかし、容量効率の高さは信頼性の低さでもある。RAID の設計においては、容量効率と信頼性はトレードオフの関係にあり、信頼性を追求すると容量効率は小さくなり、コストも増加する。よって、要求に応じた信頼性を確保しながら容量効率を高めることが重要となる。NaryRAID は、容量効率を高める技法として有効である。

disk	d_0	d_1	d_2	d_3	d_4	d_5	d_6	d_7	0	1
2^0	0	1	0	1	0	1	0	1	p_0	p_1
2^1	0	0	1	1	0	0	1	1	p_2	p_3
2^2	0	0	0	0	1	1	1	1	p_4	p_5

図 1. NaryRAID(2,3)
Figure 1. NaryRAID(2,3)

第 3 の 3 耐故障 RAID 構成法は直交 RAID である。直交 RAID は要素を共有する階層 RAID である。階層 RAID であるゆえ 3 耐故障である。加えて配線故障および RAID

コントローラの故障にも耐える。

MeshRAID44/55 はそれぞれ 2 階層の RAID4/5 で構成される。図 3 に MeshRAID44 の構成図を示す。MeshRAID444/555 はそれぞれ 3 階層の RAID4/5 で構成される。一般に RAID4/5 に基づく MeshRAID の耐故障性は階層数 n に依存し、2^n-1 で与えられる。MeshRAID の容量効率は階層 RAID に等しい。すなわち NaryRAID に劣る。各層のディスク数を N とすると容量効率は $((N-1)/N)^n$ となり、n に反比例する。

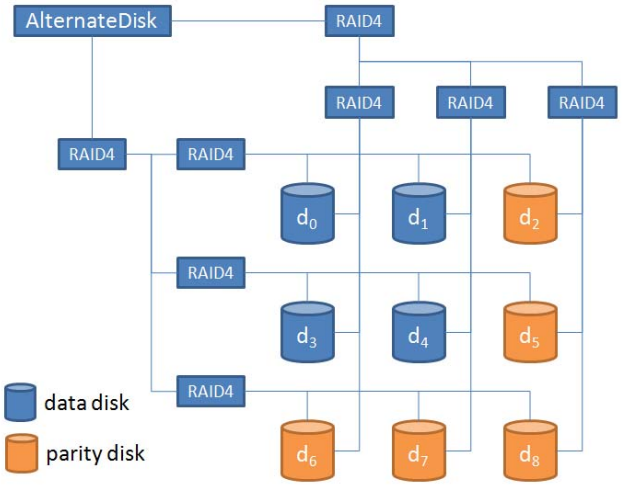

図 2. MeshRAID44
Figure 2. MeshRAID44

3. VLSD

本節では大規模ストレージ構築のための VLSD(Virtual Large Scale Disk)ツールキットについて述べる。VLSD は大規模ストレージ構築のためのツールキットであり、Java によるソフトウェア RAID 実装と NBD 実装を含む。VLSD は 100% pure Java であり、Java が動作するプラットフォームの上なら VLSD も動作する。そのため Windows や Linux が混在する環境に適している。

VLSD を用いると OS に制約されることなく NBD デバイスと RAID を自由に組み合わせることができる。最低限必要な NBD デバイスはファイルサーバの 1 つである。

Linux の nbd-server コマンドや Windows の nbdsrvr コマンドは単一ファイルを仮想ディスクとして公開する。そのため 4GB の制約がある FAT32 で動作させた場合、120GB/2GB=60 プロセスの NBD サーバを稼動させる必要がある。VLSD は複数のファイルを単一の JBOD にまとめて公開することができる。

ただし、VLSD の NBD サーバを用いた場合、ポート数の制約がある。ディスクを利用している最中はコネクションを維持するため NBD デバイスごとにポートを 1 つ消費する。ポート数はデバイス数より大きいため余裕があるが、その資源は無限ではない。数千台までは直接構成可能であるが、それを超える場合は間接的に、階層的に構成する必要がある。また、意図的に負荷を分散するために階層化することもある。この問題を解消するためにポート数に制限されない RMI を用いたディスクサーバも用意した。

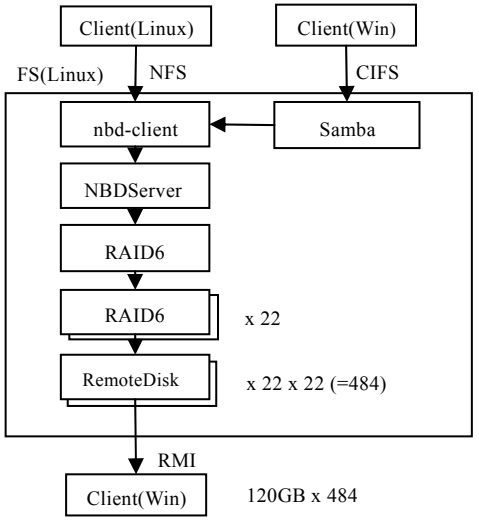

図 3. VLSD のシステム概要
Figure 3. The system overview of VLSD

図 3 に VLSD を用いて分散ストレージを構成した例を示す。クライアントは 500 台存在し、その OS は Linux または Windows である。それらはそれぞれ NFS、CIFS で 1 台のファイルサーバと通信する。クライアントは同時に NBD サーバでもある。各クライアントでは空き容量を束ねた 1 つの NBD サーバが稼動する（従来のシステムで

は複数の NBD サーバを稼動させなければならない場合があった)。ファイルサーバは Samba の稼動する Linux マシンである。ファイルサーバでは、クライアントの分だけ NBDDisk（後述）を作成し、22 の NBDDisk から 1 つずつ合計 22 の RAID6 を作成し、最後に 22 の RAID6 から 1 つの RAID6 を作成する。この RAID0 を NBD サーバで公開し、自分自身の NBD デバイスで参照する。

VLSD ツールキットには以下のクラスが含まれる。

Disk

すべての仮想ディスクのインターフェースを規定する。

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。8KiB を単位とする 1K 分木で管理する。葉ノードには 8KiB 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。ノードは必要に応じて割り当てられる。 6 階層で 8EiB-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（例えば、nbdsrvr）とも通信できる。

RemoteDisk/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。RemoteDisk 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。

WebDisk

Web サーバの資源を遠隔デバイスとして利用する仮想ディスク。WebDisk は、Web サーバで動作する REST 型 Web サービスにアクセスする。

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*=0,1,3,4,5,6)

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。

RAID4PQ/RAIDq

RAID6 と同様に 2 つのパリティ P と Q を持つ 2 耐故障 RAID である。しかし、RAID6 と異なり、RAID4 のようにパリティを専用ディスクに格納する。

RAID DP/RAIDd

水平パリティ(row parity)に加えて対角パリティ(diagonal parity)を持つ。RAID6 と同様に 2 耐故障である。2D-XOR 方式とも呼ばれる。NetApp 社の製品に使われている。2 つのパリティは互いに独立している。RAIDd を構成するにはディスク数 N は素数+1 でなければならない。

FaultDisk

耐故障性評価をおこなうためのクラス。一種のプロキシーであるが、故障を設定すると擬似的に故障を発生させる。

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．

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．

SingleRAID

SingleRAIDは単一ディスクとRAIDとして扱うアダプタである。要素ディスクが 1 つしかないRAID0とも考えられるが、すべての作業を要素ディスクへ委譲するプロキシーとして動作する点が単なるRAID0とは異なる。

StripeDisk

StripeDiskは、あるストライプグループ中から指定したブロックを抽出する。これによりRAIDの隠ぺいを迂回することができる。例えば、{D0,D1,D2}からなるRAID4に対してD1のみアクセスするには、ストライプグループを2とし、そのオフセットを1とすればよい。ただし、StripeDiskはあくまでRAIDを介して要素ディスクにアクセスする

ため、パリティディスクD2を直接読み取ることはできない。

AlternateDisk[14)]

動的に実装を切り替える。n 個の要素ディスクを持つ。要素ディスクへの要求が成功するまで順に要素ディスクを変えて試みる。すべての要素ディスクが失敗すれば全体として失敗する。

MeshRAID(44,55)

2次元直交RAIDである。直交RAIDは要素ディスクを共有する階層RAIDである。階層RAIDに比べて配線故障に強い。サブクラスにはRAID44で構成されるMeshRAID44とRAID55で構成されるMeshRAID55がある。

MeshRAID3D(444,555)[15)]

3次元直交RAIDである。直交RAID次元数は階層RAIDの階層数に該当する。次元数が増すほど耐故障性が大きくなるが、容量効率が小さくなる。サブクラスにはRAID444で構成されるMeshRAID444とRAID555で構成されるMeshRAID555がある。

4. MeshRAID MP

本論文では、複数パリティからなる直交 RAID を MeshRAID MP と呼ぶ。M は可変である。パリティ数が 1 のときは MeshRAID1P であるが、通常 1P は省略される。パリティ数が 2 のときは MeshRAID2P と呼ぶ。図 4 は最小構成の MeshRAID2P である。2P-2FT RAID の最小構成は一般に 4 台である。よって 4^2=16 台が 2 次元 MeshRAID の最小構成となる。図 3 の容量効率は必ずしもよくない。

MeshRAID2P を構成するには要素 RAID として 2P-2FT RAID を用いる。図 4 では RAID DP(RAIDd)を用いている。このような構成を MeshRAIDdd と名付ける。一般に RAIDxy と故障する階層 RAID は、下層を RAIDx、上層を RAIDy で構成する。MeshRAID では、水平階層と垂直階層が等価であるように、要素数を等しくし、x=y とする。

ここで、MeshRAIDdd の要素 RAID となる RAIDd について述べる。RAIDd の内部構成を図 5 に示す。RAIDd は専用の行パリティ R と対角パリティ D を持つ。この方式は 2D-XOR と呼ばれる。RAID6 の方式は P+Q 方式と呼ばれ、これとは異なる。2D-XOR 方式では、パリティは単純な XOR 演算のみで算出される。水平パリティ R_x は水平グループ x_i のパリティである。また、対角パリティ D_j は対角グループ x_j のパリティである。任意の 2 つのディスク故障に対して、必ず 1 つはいずれかのパリティで修復可能なディスクブロックが存在する。後は修復されたブロックを元に他のブロックを復元する。

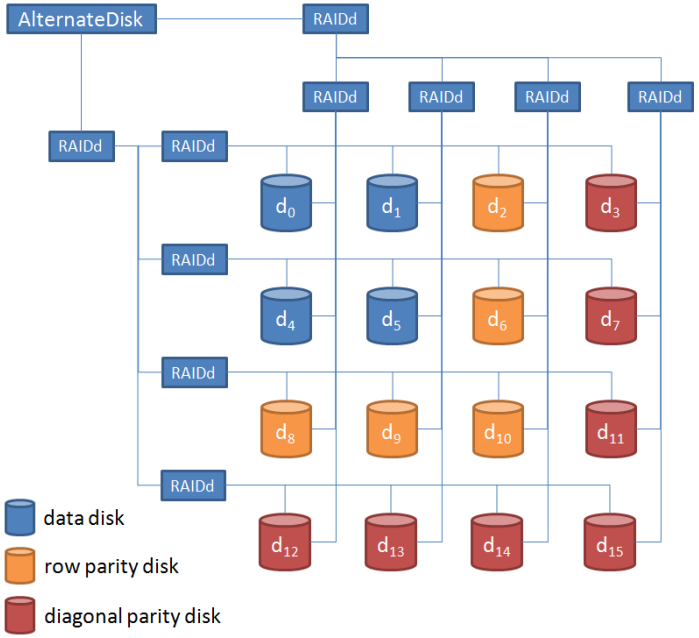

図 4. MeshRAIDdd
Figure 4. MeshRAIDdd

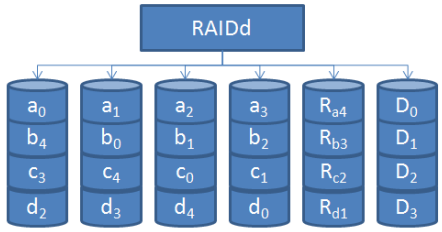

図 5. RAIDd
Figure 5. RAIDd

MeshRAID MP の耐故障性及び容量効率は階層 RAID と等しい。すなわち m パリティ RAID からなる MeshRAID *m*P のの耐故障性は$(m+1)^n$-1 である。また、その容量効率は以下の式で表される。

$$\left(\frac{k-m}{k}\right)^n = \left(1-\frac{m}{k}\right)^n \approx 1-\frac{nm}{k}$$

次に、MeshRAID2P を 3 次元化した MeshRAID3D2P と、その実例では MeshRAIDddd について述べる。MeshRAID2D には、水平階層と垂直階層の 2 つの内部階層がある。これは縦と横の 2 つの軸の順列である。同様に、MeshRAID3D には 3 つの軸の順列、すなわち 6 通りの内部階層がある。

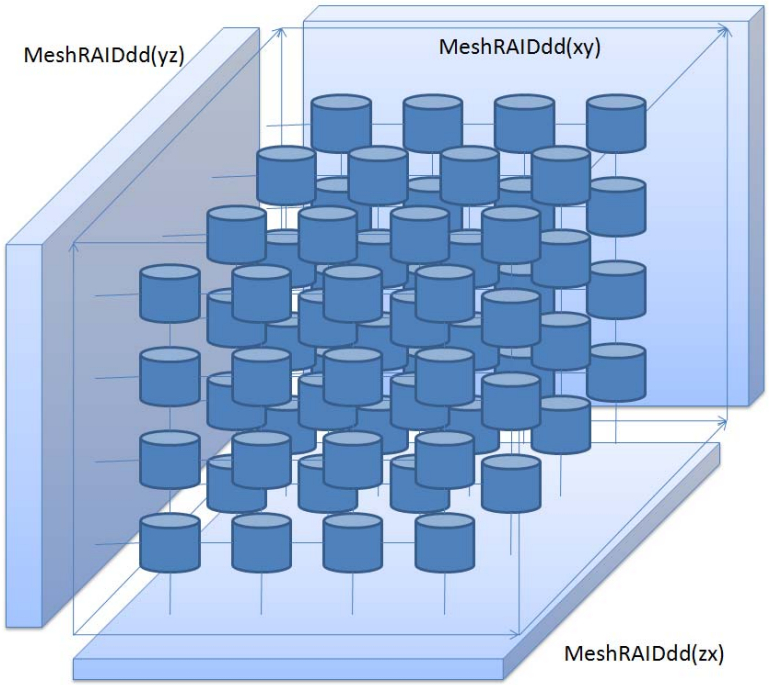

図 6. MeshRAIDddd
Figure 6. MeshRAIDddd

図 6 に MeshRAIDddd の構造を示す。各面に MeshRAIDdd が対応する。例えば、MeshRAIDdd(xy)は前面と背面にそれぞれ位置する。前面を Z,X,Y の経路とすれば、背面は Z,Y,X となる。

5. 評価

ここでは MeshRAID MP を耐故障性と性能の面から評価する。MeshRAID を次元数で分類すると MeshRAID2D, MeshRAID3D になる。MeshRAID2D は 2 次元、MeshRAID3D は 3 次元である。4 次元以上の MeshRAID は実装されていない。また、MeshRAID を要素 RAID のパリティ数 m で分類すると MeshRAID1P, MeshRAID2P になる。MeshRAID1P は 1 パリティ、MeshRAID2P は 2 パリティである。3 パリティ以上の MeshRAID は実装されていない。MeshRAID1P の評価は従来研究でなされているため、本論文では MeshRAID2P と MeshRAID2D, MeshRAID3D 組合せについて評価する。それぞれ MeshRAID2D2P, MeshRAID3D2P とする。さらに、2P-2FT RAID として RAIDd を採用する。それゆえ、それぞれの実装クラスは MeshRAIDdd, MeshRAIDddd となる。

MeshRAIDdd の理論的な耐故障性は 8 である。これは MeshRAID444 の耐故障数 7 より大きい。また、我々は 4x4=16 台からなる MeshRAIDdd 実装の耐故障性を評価した。MeshRAIDdd 実装は $_{16}C_8$ なるすべての組合せに対して修復可能であることを確認した。MeshRAIDddd の理論的な耐故障性は 26 である。また、その最小構成は 4x4x4=64 台である。残念ながら $_{64}C_{26}$ を評価するには膨大な時間がかかるため、すべての場合については確認できていない。しかし、無作為に抽出したいずれの組み合わせでも今のところ故障の発生は確認されていない。すなわち耐故障性が 26 でない反例はない。

次に、両 MeshRAID の性能を評価する。はじめに MeshRAIDdd の性能を評価する。16 台のディスクを 4x4 MeshRAIDdd, 4x4 MeshRAID44, 4x4 RAIDdd で構成し、ベンチマークテストをそれぞれ実行した。表 2 に結果を示す。比較のため表 3 に 4x4 MeshRAID44 の性能を示す。また、表 4 に RAIDdd の性能を示す。ここで SR(Small Read), SW(Small Write)は単一ブロックのアクセス、LR(Large Read), LW(Large Write)はストライプグループの全ブロックへのアクセスを表す。この評価は RAID の原典**エラー！参照元が見つかりません。**でも示されている。それぞれにおいて実行されるベンチマークは 400 回のランダムな読み書きである。評価環境は以下の通りである。OS: Windows 7 Pro 64bit SP1, CPU: Intel Core-i7 U640 1.2GHz, Mem: 2GB, SSD: 128GB。

これらの結果から、MeshRAIDdd は RAIDdd と同等の性能を持つといえる。同じ台数の MeshRAID44 より遅い。その理由は RAID44 と RAIDdd の性能差、すなわち RAID4 と RAIDd の性能差に由来する。しかし、表にも示した通り、MeshRAIDdd と RAIDdd はいずれも 8 耐故障であり、3 耐故障である MeshRAID44 よりはるかに信頼性が高い。性能を犠牲にしても信頼性を確保しなければならないとき、MeshRAIDdd は有効な選

択肢となる。

表 2. MeshRAIDdd の性能

#faults	SR[s]	SW[s]	LR[s]	LW[s]
0	0.23	0.85	0.23	0.82
1	0.23	0.82	0.22	0.81
2	0.24	0.84	0.24	0.89
3	0.23	0.78	0.24	0.81
4	0.24	0.79	0.24	0.80
5	0.24	0.80	0.24	0.78
6	0.25	0.81	0.24	0.83
7	0.25	0.86	0.24	0.85
8	0.26	0.77	0.26	0.75

表 3. MeshRAID44 の性能

#faults	SR[s]	SW[s]	LR[s]	LW[s]
0	0.13	0.57	0.12	0.52
1	0.14	0.53	0.13	0.59
2	0.13	0.54	0.13	0.50
3	0.14	0.53	0.14	0.52

表 4. RAIDdd の性能

#faults	SR[s]	SW[s]	LR[s]	LW[s]
0	0.22	0.87	0.22	0.81
1	0.24	0.85	0.23	0.84
2	0.24	0.84	0.24	0.89
3	0.23	0.80	0.24	0.79
4	0.25	0.80	0.23	0.81
5	0.24	0.81	0.24	0.78
6	0.25	0.85	0.24	0.86
7	0.25	0.86	0.24	0.85
8	0.26	0.80	0.25	0.79

次に、MeshRAIDddd の性能を評価する。64 台のディスクを 4x4x4 MeshRAIDddd, 4x4x4 MeshRAID444, 4x4x4 RAIDddd で構成し、ベンチマークテストをそれぞれ実行した。結果を図 7 に示す。これらの結果から、MeshRAIDddd は RAIDddd と同等の性能を持つといえる。しかし、同じ台数の MeshRAID444 より遅い。その理由は RAID444 と RAIDddd の性能差、すなわち RAID4 と RAIDd の性能差に由来する。図 3 の評価ではディスク番号 0 から#faults-1 までを固定的に故障させた。それゆえに必ずしも平均的な性能とはいえない。書き込みの誤差が大きいのはそのためと考えられる。

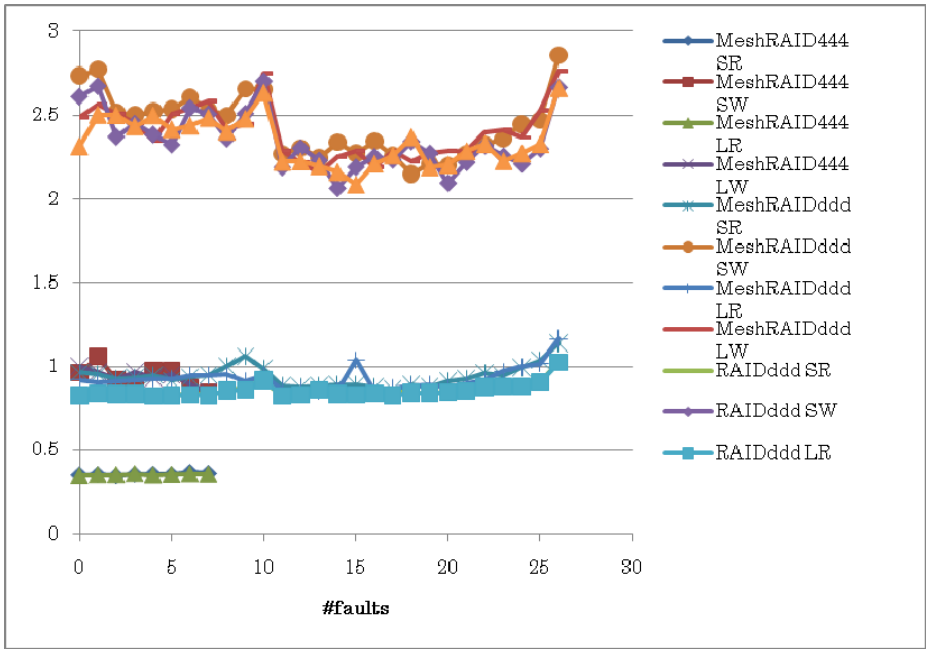

図 7. MeshRAIDddd の性能
Figure 7. The performances of MeshRAIDddd

6. まとめ

本論文では、2P-2FT RAID を要素とする直交 RAID, MeshRAID MP を提案し、その特性を解析した。また、VLSD を用いて参照実装クラス MeshRAIDdd, MeshRAIDddd を実装した。MeshRAID の耐故障性が多次元化と多パリティ化の両方で達成できることを示した。今後は、3P-3FT RAID や汎用的な mP-mFT RAID の実装を行う。

参考文献

1) Peter M. Chen, Edward K. Lee, Garth A. Gibson, Randy H. Katz, and David A. Patterson: “RAID: High-Performance, Reliable Secondary Storage,” ACM Computing Surveys, Vol. 26, No. 2, pp.145-185, June 1994
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) S. Hoon Baek, B. Wan Kim, E. Joung Joung and C. 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)
4) Erianto Chai, 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)
5) 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)
6) Minoru Uehara: "Combining N-ary RAID to RAID MP", In Proc. of 1st International Workshop on Information Technology for Innovative Services(ITIS2009) in conjunction with 2009 International Conference on Network-Based Information Systems(NBiS2009), pp.451-456, (2009.8.19-21)
7) Yuji Nakamura, Minoru Uehara: "Improving the performance of N-ary RAID by writing with XOR operation", In Proc. of 2011 25th IEEE International Conference on Advanced Information Networking and Applications (AINA2011), pp.633-638, (Biopolis, Singapore, 2011.3.22-25)
8) Minoru Uehara, Makoto Murakami, Motoi Yamagiwa: " A Proposal of 3 FT Orthogonal RAID and Its implementation in Virtual Large-Scale Disk ", IPSJ Journal Vol.52, No.2, pp.434-445, (2011.2) (in Japanese)
9) 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)
10) Minoru Uehara: "Composite RAID for Rapid Prototyping Data Grid", International Journal on Web and Grid Service, Vol.7, No.1, pp.58-73,(2011.1)
11) Minoru Uehara: "3 Faults Tolerant Orthogonal RAID for Large Storage", In Proc. of 2010 International Conference on Network-Based Information Systems(NBiS2010), pp.209-215, (2010.9.14-16, Gifu, Japan)
12) Yuji Nakamura, Minoru Uehara: "An Implementation of NaryRAID", In Proc. of 2010 24th IEEE International Conference on Advanced Information Networking and Applications (AINA2010), pp.134-141, (Perth, Australia, 2010.4.20-23)
13) Yuji Nakamura, Minoru Uehara: "Performance Evaluation of 2FT RAID", In Proc. of 3rd International Workshop on Information Technology for Innovative Services(ITIS2011) in conjunction with the 14th International Conference on Network-Based Information Systems(NBiS2011), pp.529-534, (2011.9.7-9,Tirana,Albania)
14) Minoru Uehara: "An Alternative Implementation of 3FT RAID in Virtual Large Scale Disks", INCoS2011, (TBA)
15) Minoru Uehara: "Design and Implementation of 3D MeshRAID in Virtual Large-Scale Disks", MNSA2011, (TBA)