2 耐故障 RAID の性能評価

中村祐司† 上原稔†

最近，クラウド，ライフログ，および他のアプリケーションのための大規模なオンラインストレージの需要が高まっている．以前に，我々は信頼性の高いストレージを実現するために RAID を使用して，大規模なオンラインストレージを構築するために VLSD(Virtual Large-Scale Disks)ツールキットを開発した．RAID5 は一般的に用いられるが， 1 耐故障(1FT)だけを提供するで,大規模オンラインストレージには, 適切ではない. 現在, 多くのストレージ製品が, RAID6 などの 1FT RAID 以上を使用している．しかし，1FT より高い耐性のある RAID クラスはいくつか存在する. 以前の研究で, 我々は P+Q を用いる点で RAID6 と似ている異なる 2FT RAID である RAID4PQ を開発した．しかし，それは 2 つの専用パリティを持つという点が RAID6 とは異なる．また，RAID RDP も 行パリティと対角パリティの 2 つのパリティからなる 2FT RAID である．本論文では，VLSD クラスとしてこれらの RAID タイプを実装し，そのパフォーマンスを比較する．

Performance Evaluation of 2FT RAID

YUJI NAKAMURA† MINORU UEHARA†

Recently, there has been increased demand for large-scale online storage for clouds, lifelogs, and other applications. Previously, we developed the VLSD (Virtual Large-Scale Disks) toolkit for constructing large-scale online storage, using RAID to realize reliable storage. RAID5 is commonly used, but it is not adequate for large-scale online storage because it provides only level 1 fault tolerance (1FT). Currently, many storage appliances employ more than 1FT RAID, such as RAID6. However, there are several RAID classes with greater tolerance than 1FT. In previous works, we developed another 2FT RAID, RAID4PQ, which is similar to RAID6 in the sense that it uses P+Q parities. However, it differs from RAID6 in that it has dedicated double parity. Furthermore, RAID RDP is also a 2FT RAID, in which the 2 parities are row and diagonal parities (RDP). In this paper, we implement these RAID types as VLSD classes and compare their performance.

1. はじめに

HDD とソリッドステートドライブ（SSD）のためのストレージ技術は，絶えず向上している．実際，HDD の容量は，近い将来に 5 TB に達する．SSD は容量の増加だけでなく性能も向上している．SSD は近い将来，一般的にパソコンのセカンダリストレージとして使用される．ストレージ技術におけるこれらの進歩は，クラウドベースのオンラインストレージが特に人気があり，クラウドコンピューティングの開発を推進してきた．

クラウドはスケールアウトテクノロジに基づいている．従って，クラウドベースのオンラインストレージの規模も大規模である．クラウドはデータセンターと同等であり，オンラインストレージを含め大規模なマルチプロセッサとして複数の目的に使用することができる．クラウドでは，GFS(Google File System)や Hadoop DFS(Distributed File System),などの特殊なファイルシステムがたびたび使用される. そのようなファイルシステムは大量のデータに最適化されている．そのため，通常のファイルシステムでの少量のデータには適していない．したがって，クラウドでは，NoSQL のような特殊な KVS（キーバリューストア）がファイルの代わりに使用される．従来の開発者にはファイルを使用することより KVS コードが複雑であるため, クラウドのアプリケーションを開発するのが難しくなる．VLSD (Virtual Large-Scale Disk) toolkit [4]は小規模と中規模のファイルシステムを開発者に提供することができる．我々は 2 節でこのツールキットを使用してファイルを構築する方法について説明している．

大規模ストレージにおける多数のディスクは，ストレージの信頼性がディスクの数に反比例する．RAID [1][2]はディスクの信頼性を高める技術の一つとして使用されている．RAID では RAID0-6 の 7 つのクラスがよく知られていて，RAID4-6 は容量効率と性能のバランスがよく，最も高い信頼性は RAID6 によって提供される．

しかし，RAID6 は RAID5 の拡張なので RAID5 の特徴を継承し，分散パリティを使用している，分散パリティは通常のハードウェア RAID のボトルネックを回避するのに有効である．しかし，ソフトウェア RAID ではキャッシュがうまく動作しない．そのため, 我々は RAID6 のように P と Q の 2 つのパリティによる RAID4PQ[5]を開発した．しかし，これらのパリティはパリティ専用ディスクに格納される．RAID4PQ という名前は RAID4 のブロックストライピングと, RAID6 の P+Q パリティを示している．RAID4PQ も RAID6 と同様に 2FT を提供する．

本論文では，VLSD クラスとして別の 2FTRAID である RAID RDP(row and diagonal

† 所属
東洋大学大学院情報システム専攻
Dept. of Open Information Systems Graduate School of Engineering Toyo University, Japan

parity)を提示する．RAID4PQ と RAID6 は行パリティのみを持っているが RAID RDP は行パリティと対角パリティを持っている．行パリティでは，パリティはストライピンググループから計算される．対して，対角パリティではいわゆる RDP(row-diagonal parity)アルゴリズム[6]によって全て異なるストライピンググループから計算される．また，RAID RDP，RAID4PQ，RAID6 による異なる VLSD 実装の比較も行った．

本論文の構成は以下の通りである．2 節では，VLSD を紹介し，3 節では RAID4PQ を説明する．4 節と 5 節では RAID RDP の設計と実装を示す．6 節では，RAID RDP の性能を RAID6 と RAID4PQ と比較する．最後に，結論を述べる．

2. 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[4]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ーの 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 サーバー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 サーバーを用いた場合，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ーも用意した．

図 1 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ーと通信する．クライアントは同時に NBD サーバーでもある．各クライアントでは空き容量を束ねた 1 つの NBD サーバーが稼動する（従来のシステムでは複数の NBD サーバーを稼動させなければならない場合があった）．ファイルサーバーは Samba の稼動する Linux マシンである．ファイルサーバーでは，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 RAID0File を NBD サーバーで公開し，自分自身の NBD デバイスで参照する．

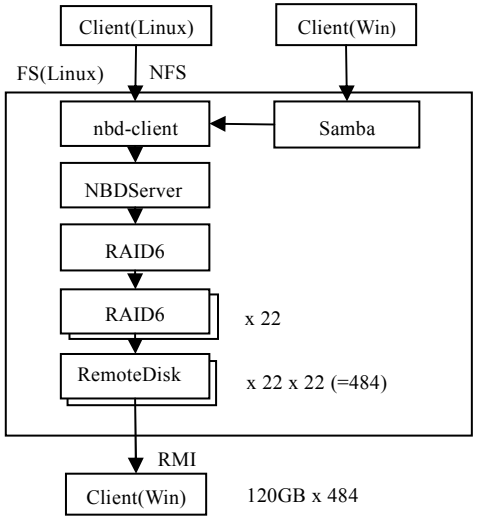

図 1　VLSD のシステム概要
Figure 1　The system overview of VLSD

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．

NamedFolderDisk

ページを同一ディレクトリの異なるファイルに分割して保存する．アクセスは非常に遅い．ページサイズを大きくすれば，容量を大きくできるが，ページ単位アクセスが遅くなる． また，ファイルシステムによっては，ページ数が一定数を超えるとディレクトリのシークが極端に遅くなる． JBOD+FileDisk と基本的に等しい．

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 サーバー実装とも通信できる．

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 に対応するサーバーは 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 に対応するサーバーは 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. RAID4PQ

RAID4PQ[5]は RAID6 の分散パリティディスクの変わりに 2 つの専用パリティディスク P と Q を用いた RAID クラスである．7 つの RAID クラスは必ずしも直行した概念で構成されていない. 例えば, RAID4,5,6 はブロック単位ストライピングであるが，RAID4 は専用パリティディスクで，RAID5,6 は分散パリティとなっている．また，パリティディスクの数は RAID4,5 では 1 つで, RAID6 は 2 つとなっている. したがって，2 つの専用パリティの RAID クラスは存在しない．それは，ハードウェア RAID において，RAID5 が RAID4 より優れていることが知られているので，そのような RAID クラスが必要ないと考えられるためである. しかし, ソフトウェア RAID においては，キャッシュが働くので，このようなクラスにも有用性が認められる．ハードウェア RAID では，キャッシュは無視される．さらに，我々は以前の研究で，MeshRAID と呼ばれる，新たな直交 RAID を提案した[8]．分散パリティは，このような直交 RAID には適していない．代わりに，専用パリティは容易に理解できるので複雑な RAID により適していると言える．そこで我々は，VLSD に 2 つの専用パリティディスクを持つ 2FT RAID クラスである RAID4PQ を実装した．VLSD では RAID6 がすでに実装されているので，RAID4PQ は RAID6 のサブクラスとして，ストライピンググループにおけるパリティディスクの場所を特定するメソッドを修正して実装した. VLSD では，RAID はソフトウェアコンポーネントであり，容易に拡張可能である．

文献[5]では，RAID4PQ の性能は，キャッシュが使用されていなくても RAID6 とほぼ同等であることが示されている．パリティディスクの割り当ての違いは，OS によって隠されている．RAID 6 の場合，パリティディスクとして使用されている場合は全てのディスクがキャッシュされる必要があるため，キャッシュはあまり効果的ではない．しかし，RAID4PQ の場合では，2 個のディスクだけがキャッシュされる必要があるので，キャッシュは有効である．

RAID4PQ では，ディスクアクセスがパリティディスクに集中され，これはしばしばボトルネックを引き起こす． しかし，キャッシュが使用される場合，RAID4PQ にはボトルネックはなくなる．

4. RAID RDP

RAID RDP は RDP メソッドに基づく 2FT RAID クラスである．RAID RDP は行パリティと対角パリティの 2 つのパリティを持つ．RAID RDP では，RAID6 とは異なりパリティは排他的論理和演算のみを用いて計算される．各データブロックは行とパリティの両方のパリティグループに属している．RAID RDP を構築するためにはディスクの数は 2 よりも大きい素数 p に対して， p +1 である必要があり，各ディスクは p -1 個のストライプに属する．

図 2 はディスク 6 台で 4 つのストライプで構成される p＝5 の RAID RDP を示している．ここではブロックの色は，ブロックが属する対角パリティのグループを示して

いる．対角パリティディスクは全てのグループのパリティを格納しない．しかし，各パリティは少なくとも 1 つの行または対角パリティディスクに格納する必要がある．

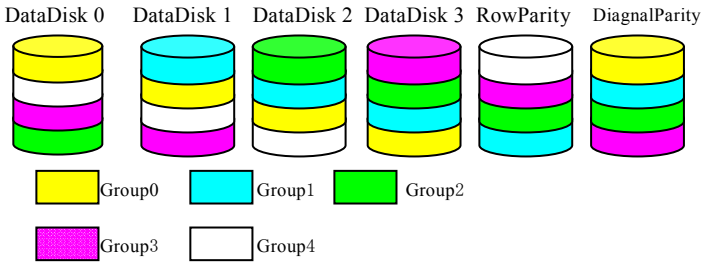

図 2　RAID RDP の構造
Figure 2　Structure of RAID RDP

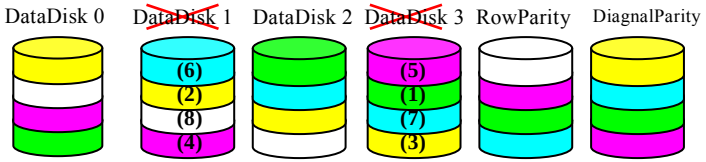

図 3　RAID RDP での修復
Figure 3　Repair in RAID RDP

RAID RDP は次のように修復される．例として，図 3 ではデータディスク 1 と 3 の故障を仮定している．この場合，2 つの故障が各行パリティグループの中にある．よって，行パリティだけを用いて修復することはできない．しかし，対角グループのうち少なくとも 2 つ修復可能なグループがある．1 つのディスクの故障だけを含んでいるグループであれば修復は可能である．このケースでは対角グループ 2 と 4 が修復可能となる．しかし，対角パリティディスクにはグループ 4 のブロックは存在しない．そこで，はじめにディスク 3 のグループ 2 であるブロック 2 から修復が行われる．

図 3 の例に基づいて修復は以下のように行われる．

(1) はじめに，対角パリティグループ 2 に 1 つだけの故障があるのでディスク 3 のブロック 2 が対角パリティを用いて修復される．

(2) 次に，行パリティグループに単一の故障があるので，ディスク 1 のブロック 2 が行パリティを用いて修復できる．

(3) 次に，対角パリティグループ 0 に単一故障ができるので，ディスク 3 のブロック 4 が対角パリティから修復される．

(4) 次に，行パリティグループに単一の故障があるので，ディスク 1 のブロック 4 が行パリティを用いて修復できる．

(5) 次に，対角パリティグループ 3 に単一故障ができるので，ディスク 3 のブロック 1 が対角パリティから修復される．

(6) 次に，行パリティグループに単一の故障があるので，ディスク 1 のブロック 1 が行パリティを用いて修復できる．

(7) 次に，対角パリティグループ 1 に単一故障ができるので，ディスク 3 のブロック 3 が対角パリティから修復される．

(8) 最後に，行パリティグループに単一の故障があるので，ディスク 1 のブロック 3 が行パリティを用いて修復できる．

5. RAIDRDP の実装

VLSD に RAID DP のクラスを追加し実装を行った．RAID RDP は 1 台までの故障時には RAID4 と同じ振る舞いをする．よって，VLSD の RAID4 クラスを改良し実装している．ライトでは書き込みブロックに対応する対角パリティの更新の更新を追加する．リードでは故障ディスクが 2 台のときの対角パリティを用いた読み取りを行う．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．

RAID DP

クラス RAID DP は RAID のサブクラスで RAID RDP の実装である．RAID4 構成のブロック単位ストライピングと行パリティ専用ディスクの作成に加え，データディスクと行パリティ専用ディスクから，ブロック単位対角グループと対角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対角パリティディスク，後ろから 2 番目のディスク番号のものを行パリティディスクとして構築する. RAID DP クラスは, 4 節で与えられたアルゴリズムを実装するために，次の 3 つのメソッドを持つ．

seekInSelectedDisk(long pos)

このメソッドは long 型のパラメータ pos を取り，pos の位置にあるデータを持ったディスクを特定し，ディスク上の行位置を求め，最後にその行位置を返す．

seekInSelectedDisk(long DPgroup, int no)

このメソッド long 型のパラメータ *DPgroup* と整数 *no* を受け取り，*no* によってディスク番号を選択し，パリティディスク上の *DPgroup* の位置を求め，最後のその位置を返す．

getDiagonalGroupNumber(int no, long stripe)

このメソッドは Disk 番号である *no* の行ブロック番号 *stripe* が属するパリティグループを特定し，対角パリティグループ番号 *DPgroup* を返す．

RAID RDP は排他的論理和書き込み[7]による性能改善技術を適用している．排他的論理和書き込み方式は RemoteDisk と DiskServer 間のトラフィックを減らすことができる. 図 4 の例は, 従来方式の RAID の書き込みシーケンスを表している. 図 4 では, XOR 演算はクライアント側で 2 度実行されている. これにより不要な読み込みが発生する．これをサーバー側で XOR 演算を実行することにより，不要な読み込みを減らしている．この手法を”Write XOR”とする．図 5 に改良されたシーケンスを示す．このメソッドは読み込み操作の回数を減らし，特にインターネットのような大規模な潜在的なネットワークに効果的である．

表 1 は一般的な RAID の読み書き回数を示している．提案手法は理論的に従来手法より常に優れている．提案手法の有用性は XOR に基づくパリティの数に依存する．しかし，この手法は XOR 演算をサーバー側で実行する必要があり，これはハードウェア RAID で実行することは難しい．従ってこの方法は VLSD のようなソフトウェア RAID に適している．

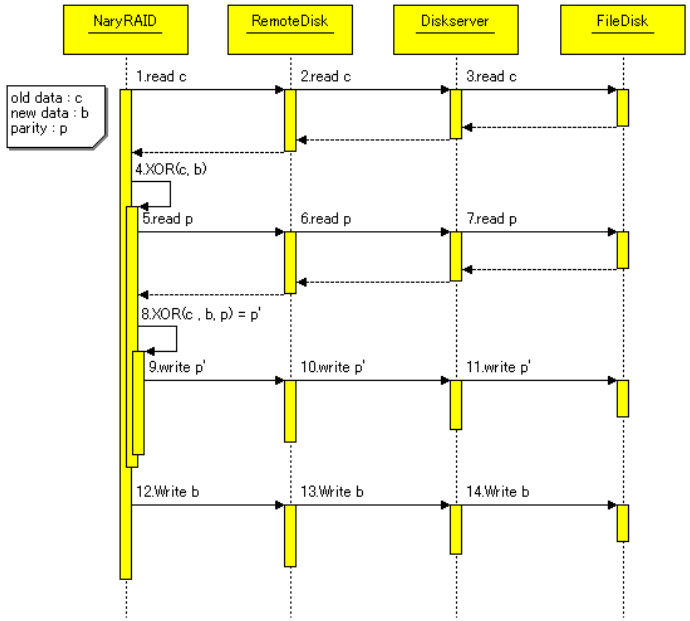

図 4　NaryRAID の書き込みシーケンス
Figure 4　Writing sequence of NaryRAID

表 1　一般的な RAID の読み書き回数
Table 1　Number of Read/Writes in general RAIDS

	# read	# write
NaryRAID(N,n)	n+1	n+1
NaryRAIDWriteXOR(N,n)	1	n+1
RAID4, 5	2	2
RAID4, 5 WriteXOR	1	2
RAID6, 4PQ, RDP	3	3
RAID6, 4PQ, RDP WriteXOR	1	3

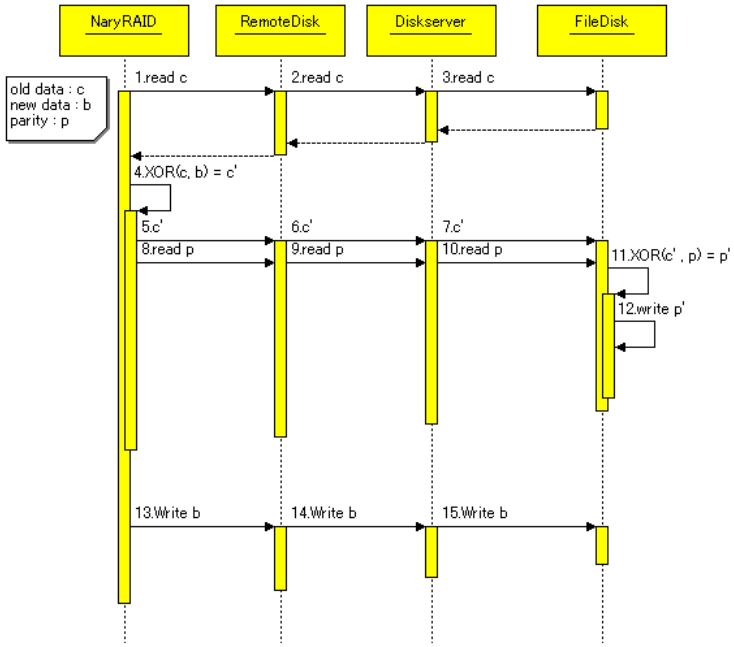

図 5　NaryRAID WriteXOR の書き込みシーケンス
Figure 5　Writing sequence in NaryRAID WriteXOR

6. 評価

ここでは，RAID RDP の性能を評価し，RAID6，RAID4PQ と RAID RDP を含む 2FT RAID の特徴を比較する．

初めに，ディスクサイズに関する性能評価を行った．結果を図 6 に示す．この実験ではディスク台数を 6 台としている．この場合 RAID6，RAID4PQ，RAID RDP のような 2FT RAID では 4 台のデータディスクと 2 台のパリティディスクとなり，RAID4 のような 1FT RAID では 5 台のデータディスクと 1 台のパリティディスクとなる．1 つのディスクサイズは 1MB，10MB，100MB，1GB と変えて実験を行った．nF の表記は故障の台数が n 台であること示している．ベンチマークでは 400 のランダムアクセスを行っている．ここでは SmallRead(SR)，SmallWrite(SW)は同一ディスクに対するブロックサイズのアクセスを表し，LargeRead (LR)，LargeWrite(LW)はグループ全体におよぶストライピングサイズのアクセスを表す．これらは元の RAID の論文での RAID の一般的な評価基準として定義されている．

実験の結果から，SR ではどの RAID も差がなく故障時の傾向も同じであるといえる．SW では，RAID RDP は常に RAID6 よりも早いが故障時には RAID4PQ よりも遅くなっている．性能はディスクサイズに依存し，ディスクサイズがより大きいほど，その差はより大きくなる．LR では，RAID RDP は全てのクラスと比べてもっとも遅い．LW では RAID RDP は RAID4PQ よりも遅いが，RAID6 よりは高速である．しかし，これらの結果は VLSD のようなソフトウェア RAID において得られたものであり，ハードウェア RAID においては結果が異なる場合がある．

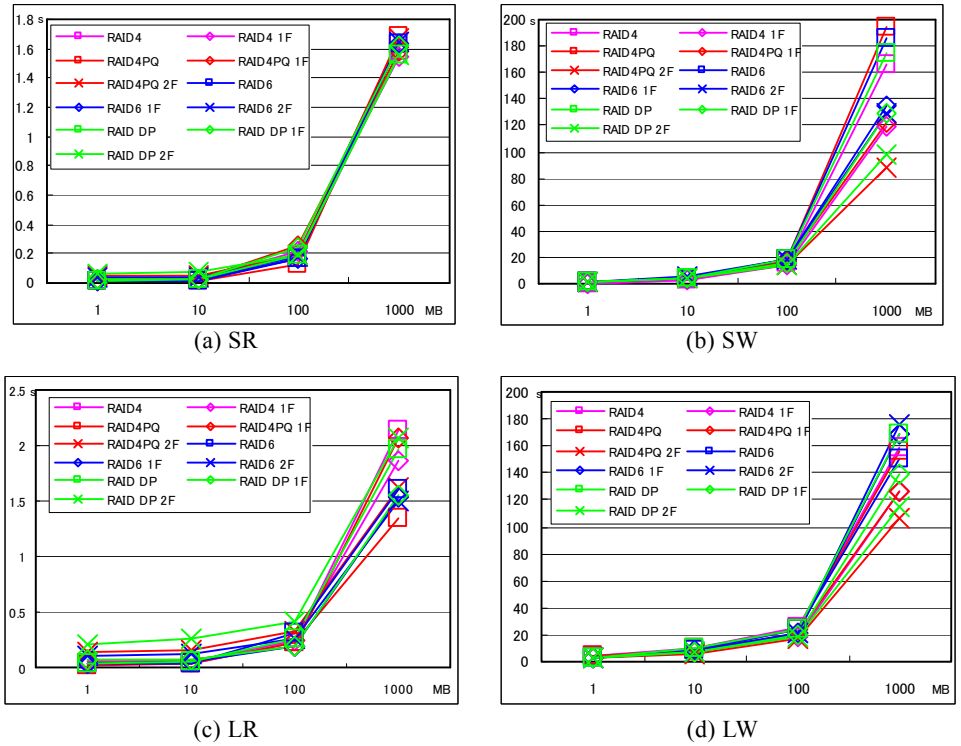

図 5　2FT RAID の性能比較
Figure 5　Performance comparison of 2FT RAIDs

7. まとめ

本論文では, 我々は RDP 方式に基づく 2FT RAID として RAID RDP を提案し, VLSD を用いて RAID RDP クラスを実装した．そして他の 2FT RAID と性能の比較評価を行った. 評価では, RAID RDP と RAID6, RAID4PQ,そして RAID4 と比較を行った. RAID4 は 1FT RAID ではあるが他の RAID が RAID4 に基づいているので 2FT RAID を RAID4 と比較することには意味がある．また，我々は故障時の性能評価も行った．結果として読み込み性能はほとんど同じであることを示している．書き込み性能では性能はディスクサイズに依存していて，ディスクサイズが小さいときはあまり差がないが，ディスクサイズを大きくすると RAID RDP は RAID4PQ よりやや遅く，RAID6 よりやや早い傾向になる．全体的に RAID RDP は他の 2 耐故障に対して大きく劣ることはなく実際の運用にも利用できると考えられる．

今後の課題としては，我々は今回実装した RAID RDP を NaryRAID と組み合わせて利用することを考えている．文献[6]では，我々は NaryRAID と RAID6，RAID4PQ または RAID RDP のような他のマルチパリティ RAID を組み合わせた NaryRAID MP を提案している．これらの実装を行う．そして，実際にいくつかのマシンを用いたシステムの性能評価を行う．

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] 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, June 1994, pp. 145–185

[3] 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)

[4] E, Chai, M, Uehara, H, Mori, N, Sato: “Virtual Large-Scale Disk System for PC-Room”, LNCS 4658, Network-Based Information Systems, pp.476-485, (2007.9.3-4)

[5] Minoru Uehara: "RAID4PQ: P+Q based RAID with dedicated parity disks", In Technical Reports of 2010 Research Institute of Industrial Technology, Toyo Univ., pp.138-139, (2011.2.25) (in Japanese)

[6] Peter Corbett , Bob English , Atul Goel , Tomislav Grcanac , Steven Kleiman , James Leong , Sunitha Sankar, Awarded Best Paper! “Row-Diagonal Parity for Double Disk Failure Correction”, Proceedings of the 3rd USENIX Conference on File and Storage Technologies, March 31-31, 2004, San Francisco, CA

[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: "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)