複数ディスクからなるストレージシステムの省電力化手法における電力削減効果の比較および評価

引田　諭之† 　LeHieu Hanh† 　横田　治夫†

† 東京工業大学大学院情報理工学研究科計算工学専攻　〒 152–8552 東京都目黒区大岡山 2–12–1
E-mail: †{hikida,hanhlh}@de.cs.titech.ac.jp, ††yokota@cs.titech.ac.jp

あらまし　ストレージシステムの省電力化は重要な課題の一つである．我々はこれまでにプライマリ・バックアップ構成を有効活用した省電力化手法として RAPoSDA を提案し，概算式を用いてその省電力効果を確認してきたが，ワークロード下での効果は検証出来ていなかった．本研究では，新たに構築したシミュレータを用いてワークロードを与えた実験を行い RAPoSDA および関連手法の電力削減率と性能に関して比較評価を行う．

キーワード　ストレージ，省電力，シミュレーション

An Evaluation and Comparison of Power Reduction for Multiple Disk Drive Storage System

Satoshi HIKIDA†, Le HIEU HANH†, and Haruo YOKOTA†

† Department of Computer Science, Graduate School of Information Science and Engineering,
Tokyo Institute of Technology　2–12–1 Ookayama, Meguro-ku, Tokyo 152–8552, Japan
E-mail: †{hikida,hanhlh}@de.cs.titech.ac.jp, ††yokota@cs.titech.ac.jp

Abstract　Nowadays, Reducing the power of storage system has been an important issue as data reliability. Thus we have proposed *RAPoSDA*(Replica Assisted Power Saving Disk Array) that method reduce the power consumption of the storage systems with keeping the reliability of data by utilizing the primary-backup configuration. But we have not verified its performance and efficiency of power reduction under workloads yet. In this paper, we evaluate and compare the performance and efficiency of power reduction of RAPoSDA and related method by the simulator we developed.

Key words　Storage, Power Reduction, Simulation

1. はじめに

ストレージの省電力化は近年重要な課題として認識されている．情報技術の発達によるデータ量の爆発的な増大によって，必要とされるストレージの容量も増加し，それに伴いストレージ装置も大規模化しており，データセンター等においては今後ストレージが消費電力量の多くを占めるだろうと予想される．

我々はこれまでに，プライマリ・バックアップ構成を有効活用したストレージの省電力化手法である RAPoSDA(Replica Assisted Power Saving Disk Array) を提案してきた [10]．RAPoSDA は，キャッシュメモリとディスクドライブの双方でプライマリ・バックアップ構成をとることでストレージの信頼性を確保し，個々のディスクの回転状況を考慮してディスクへ書き込むデータを決定し，不必要なディスクのスピンアップを回避することで消費電力を抑制する手法である．

文献 [10] では，消費電力量を見積もる概算式を用いて RAPoSDA および関連手法である MAID [1] の電力削減率について比較し，RAPoSDA がより省電力効果の高いことを示した．しかし概算式では，アクセスが時間的に変化するワークロード下における消費電力量や性能に関しては検証できていなかった．

ワークロードを用いた評価方法としては実機による実験とシミュレーションプログラム（シミュレータ）によるシミュレーション実験という選択肢があるが，実機実験ではストレージの構成変更や大規模環境の構築が容易ではないため，今回はシミュレータによるシミュレーション実験をおこなうこととする．

シミュレーション実験に際し，我々は新規にストレージのシミュレータを構築し，そのシミュレータを用いて検証をおこなう．与えるワークロードは，Zipf 分布に基づくデータアクセスの偏りを持つ人工的なものを用い，読み出し比率が異なるいくつかのワークロード下での RAPoSDA や関連手法における省電力効果や性能について比較評価を行う．

なお，本論文の構成は以下の通りである．2. 節では今回比較

表 1　ディスクドライブの状態と消費電力
Table 1　A table of status and corresponding power consumption

状態	I/O 処理	RPM	ヘッド位置	消費電力
Active	処理中	最高回転	ディスク上	大
Idle	処理なし	最高回転	ディスク上	中
Standby	処理なし	0	ディスク外	小

評価を行う省電力化手法の説明を行い，3. 節でシミュレータの概要を述べ，4. 節でシミュレーションにおけるストレージの構成や用いるワークロードの詳細を説明し，シミュレーション結果についても考察を行う．5. 節では関連研究ついて述べ，最後の 6. 節でまとめと今後の課題について述べる．

2.　評価対象の省電力化手法

本節では，シミュレータを用いて評価を行う省電力化手法について説明する．まず最初にストレージの主要な構成要素であるディスクドライブの消費電力について説明し，次に評価対象である RAPoSDA と MAID について概要を述べる．

2.1　ディスクドライブの消費電力

ディスクドライブは，データを記録するプラッター（円盤）や，プラッターを回転させるスピンドル，データの読み取り/書き込みを行うヘッドとヘッドをディスク上で移動させるアーム等の機械部品と，ディスクドライブの動作を制御する制御部品で構成されている．大部分の電力は機械部品であるスピンドルやアームによって消費されるが，それら機械部品の動作状況に応じて大きく 3 つの状態の状態遷移としてディスクドライブの消費電力はモデル化できる（表 1）．

表 1 より，三つの状態中で一番消費電力が大きいのは Active 状態であり，一番小さいのは Standby 状態である．また，ディスクの回転開始（Spin-up）時および回転停止（Spin-down）時にも一時的であるが大きな電力が消費され，特に回転開始時には Active 状態よりも大きくなることがある．このような性質から，ストレージの省電力化のためにはアクセスの無いディスクドライブは出来るだけ長時間に渡って回転停止させて Standby 状態を長く保つことが重要であるが，無闇に回転を停止させるだけではディスクアクセスが発生するたびに回転開始が必要になり，消費電力は従来よりもかえって増大してしまうおそれもある．

そこで，ディスクのスピンダウンはどのような基準で行うべきかを判断するためにブレイクイーブン時間（break-even time）というものが用いられる．ブレイクイーブン時間とは，アイドル状態に対し，スタンバイ状態で節約できるエネルギーと，スピンアップとスピンダウンとで消費されるエネルギーの合計が等しくなる時間のことをいう．もしスタンバイ状態の期間がこのブレイクイーブン時間よりも長い場合，その分だけ省電力効果がある．

2.2　RAPoSDA について

RAPoSDA [10] は，多数のディスクドライブを組み合わせたストレージシステムの省電力化を対象としている．データセンター等で実際に運用されるときは信頼性の確保が重要であるため，データは冗長化されて複数のディスクに保存されることが多い．RAPoSDA ではデータの信頼性を確保するために，キャッシュメモリとディスクドライブの双方でプライマリ・バックアップ構成をとるようにし，データ配置方法やディスクへの書き込みのタイミングを工夫することで省電力化を実現している．ディスクドライブにおけるデータの配置方法は Chained Declustering [4] を用いている．

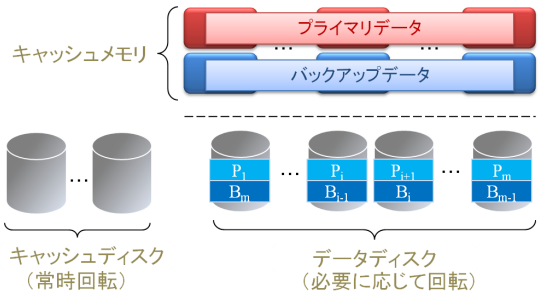

図 1　RAPoSDA の全体構成図
Fig. 1　Configuration of RAPoSDA

2.2.1　構　　成

RAPoSDA は主に次の要素により構成される（図 1）．

- キャッシュメモリ
- キャッシュディスク
- データディスク

キャッシュメモリ：キャッシュメモリには揮発性の RAM を用いることを前提としており，信頼性を持たせるために異なる二つのキャッシュメモリにプライマリとバックアップのデータを持たせる冗長構成をとる．そのため，あるキャッシュメモリは二つの領域（プライマリ層およびバックアップ層）でデータを保持することになる．それぞれのキャッシュメモリは個別の電源系統を持ち，UPS（無停電電源装置）等で断電対策が施されているものとする．1 つのキャッシュメモリは 1 つ以上のディスクドライブで共有される．

キャッシュディスク：データをキャッシュするためのディスクである．キャッシュメモリでは容量に限界があることと，多くのワークロードでは読み出し要求が書き込み要求よりも多いという状況から，キャッシュメモリは書き込みデータのバッファを主目的とし，キャッシュディスクは読み出し専用とする．読み出し要求に迅速に対応するキャッシュディスク数は後述するデータディスク数よりも少ない構成とし，常時回転させておくことでディスクのスピンアップに伴う応答遅延を回避する．

データディスク：実際のデータを格納するディスクドライブである．キャッシュメモリ上のバッファがあふれた場合や，読み出し時にキャッシュミスした時などにディスクへのアクセスが発生する．ディスクアクセスがある閾値時間を超えて発生しなかった場合はそのディスクドライブをスピンダウンさせてスタンバイ状態に移行する．このスタンバイ状態期間がブレイクイーブン時間よりも長ければ長いほど省電力効果が得られる．また，データディスクでも信頼性の確保のためにプライマリ・

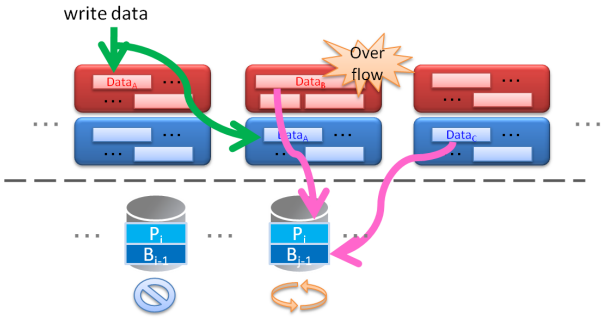

図 2　RAPoSDA 書き込み処理
Fig. 2　Write process of RAPoSDA

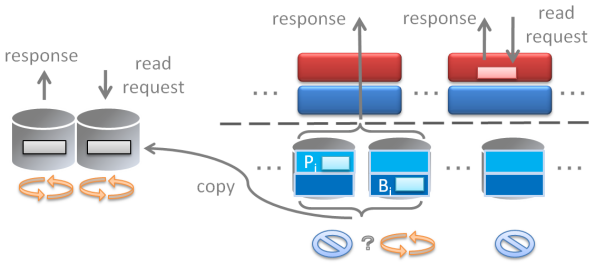

図 3　RAPoSDA 読み出し処理
Fig. 3　Read process of RAPoSDA

バックアップ構成をとる．

2.2.2　動　　作

書き込み

データの書き込みは，はじめにキャッシュメモリに書き込む．その際，データは電源系統の異なる別々のキャッシュメモリ上のプライマリ領域とバックアップ領域に書き込まれる．キャッシュメモリのバッファ容量の閾値を超える場合に，データディスクへバッファ中のキャッシュデータを書き込む（図 2）．

あるデータをキャッシュメモリに書き込んだ際にバッファ容量の閾値を超えた場合，そのデータに対応するデータディスクに対しディスクの回転状況を確認し，もし回転停止（スタンバイ状態）のときはそのディスクドライブをスピンアップさせる．その後，バッファ容量の閾値を超えたキャッシュメモリ上の全データに対し，対応するデータディスクへ書き込みを試みる．但し，実際にデータディスクへデータを書き込むのは該当データディスクが回転中の場合のみに限定する．このとき，少なくとも一つのデータディスクは回転中であるので，バッファ上の全てのデータは書き込みきれなくともバッファ容量に空き領域を確保することは可能である．

あるデータディスクが回転中であり，バッファ上のデータを書き込む際に，その書き込みデータがプライマリ領域データだった場合，同一ディスクのバックアップ領域データに対応する別のキャッシュメモリ上のデータもこのタイミングで同時に書き込む．バックアップ領域データは別のキャッシュメモリ上に存在しているため，バッファ容量の閾値は超えていない可能性もあるが，このタイミングで書き込むことによってデータディスクをスピンアップさせる回数とディスクアクセス頻度の抑制を図っている．なお，書き込むデータがバックアップ領域データだった場合も同様に，同一ディスクのプライマリ領域に対応するデータを同時に書き込むようにする．

データディスクに書き込まれたデータはキャッシュメモリ上から削除し，さらに今後の読み出し処理に対応するためにキャッシュディスクにデータをコピーしておく．

読み出し

読み出し処理は，キャッシュメモリ，キャッシュディスクの順序でキャッシュデータを確認し，該当データが見つかった時点で応答を返す．キャッシュメモリやキャッシュディスクにもデータが見つからなかった場合，データディスクから該当データを読み出す（図 3）．

キャッシュメモリはプライマリ層とバックアップ層に分かれているので，該当データがどちらかの層に存在していればデータはキャッシュメモリから読み出される．

キャッシュメモリ上に該当データが存在しなかった場合は，キャッシュディスクにデータが存在するかを確認する．もしキャッシュディスク上にデータが存在していれば，データはキャッシュディスクから読み出される．

キャッシュメモリおよびキャッシュディスクに該当データが存在しなかった場合，データディスクから読み出す必要があるが，データディスクはプライマリ・バックアップ構成をとっているのでどちらのディスクから読み出すかを決定する必要がある．以下に対象データディスクを選択するパターンを示す．

- **片方のみ回転中** 回転している方のディスクから読み出す
- **両方回転中** バッファ容量の多い方のディスクから読み出す
- **両方停止中** 停止期間が長い方のディスクから読み出す

最後に，データディスクから読み出したデータは，今後の読み出し処理に備えてキャッシュディスクにコピーされる．

2.3　MAID について

MAID [1] はキャッシュディスクを用いたストレージの省電力化手法である．データアクセスの局所性に着目し，頻繁にアクセスされるデータをキャッシュディスクと呼ばれる常時回転中で小数のディスクドライブに集約し，その他の大多数のディスクドライブへのアクセスを抑制する．アクセスのないアイドル時間がある閾値を超えたディスクドライブはスピンダウンさせてスタンバイ状態にする．

書き込み要求も読み出し要求も最初にキャッシュディスクで受け付けているため，キャッシュディスクの容量やデータ転送能力が全体性能のボトルネックとなるおそれがある．

また，MAID で目標としているのが多少の性能劣化は許容してでも大幅な省電力化を実現することであるが，ストレージの重要な要素である信頼性については考慮されていない．

3.　シミュレータ

本節では RAPoSDA および MAID を評価するために今回新規に構築したシミュレータについて説明する．

ストレージシステムの性能を評価するシミュレータは既にいくつか存在しており，その中でも DiskSim [2] は様々な研究で

利用されている．しかし DiskSim は消費電力の測定までは考慮されておらず，そのままでは今回の比較評価に用いることができない．Dempsey [9] は DiskSim を拡張し，ストレージの消費電力もシミュレート出来るようにしているが，あらかじめ実際のディスクドライブの電力を測定しておく必要があり，シミュレーションに用いるハードディスクのモデルが制限されてしまうという問題がある．

そのため，我々は新規にストレージのシミュレータを構築し，そのシミュレータ上で RAPoSDA や関連手法である MAID を評価することとした．なお，新たに構築したシミュレータは一部のコンポーネントを差し替えるだけで柔軟にその構成を変更できるため，様々な手法間の比較評価が容易に行えるという特徴を持つ．

3.1 概　　要

シミュレータは，シミュレーション環境上に構築したストレージサブシステムに対し時間的に変化するワークロードを与え，応答時間やスループット等の性能や各ディスクドライブで消費された電力をシミュレートする．ワークロードは人工的に生成したものや外部で公開されているトレース等をベースにした実運用環境のワークロードを用いる．

3.2 構　　成

シミュレータの全体構成を図 4 に示す．このシミュレータは五つのコンポーネントで構成されており，以下で各コンポーネントの概要を説明する．

Workload Generation：パラメータを基に人工的なワークロードを生成する．人工的なワークロードは Zipf 分布に従うアクセス分布とポワソン到着に従うアクセス到着率を持つワークロードを発生させる．また，サーバー等のトレースをパラメータとして与えた場合，そのトレースベースのワークロードを生成する．

Data Layout Management：データ管理の中心的なコンポーネントであり，このコンポーネントによって RAPoSDA や MAID 等の各種省電力化手法が実現される．

Storage Devices：各種デバイス（メモリおよびハードディスクドライブ）の動作をシミュレートする．

Log Collection：各デバイスにおける情報を収集し，ログとしてテキストファイルに出力するコンポーネントである．ログとして収集する情報は，各リクエストの応答時間や，各ディスクドライブの消費電力，キャッシュのヒット率，ディスクアクセス時のディスクが回転している割合等である．

Analysis：Log Collection で出力された各種ログファイルを解析し，平均応答時間やスループット，消費電力等を算出する．

3.3 シミュレータの動作

シミュレータの起動時には，ワークロード，シミュレータの設定値，各種デバイス（キャッシュメモリ，キャッシュディスク，データディスク）のモデル情報をパラメータとして渡す．その後シミュレーション内のクライアントがワークロードで指定された時刻でリクエストを生成し，StorageManager にリクエストを送信する．StorageManager は DataLayoutManager にリクエスト情報を渡し，DataLayoutManager がデータの配置先を決定して各デバイスへのアクセスを指示する．各デバイスの動作状態に関する情報はログとして LogCollector に収集される．シミュレータの実行完了後に Analyser によってログを解析する．

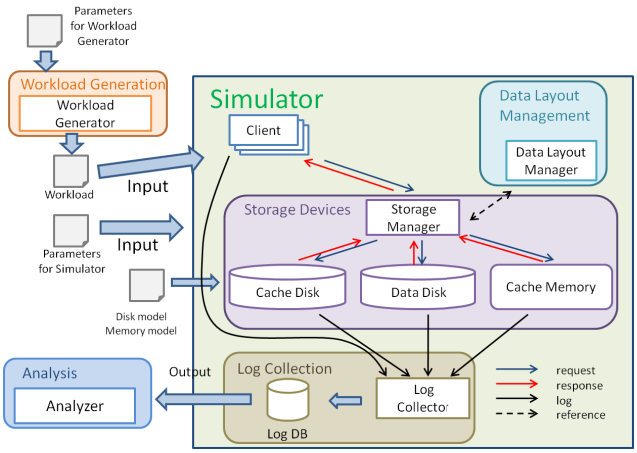

図 4　シミュレータの全体構成
Fig. 4　Configuration of the simulator

表 2　シミュレーションで用いる HDD のパラメータ
Table 2　Prameters of hard disk drive using by the simulator

parameter	value
容量（TB）	2
プラッター数	5
ディスク回転数（RPM）	7200
ディスクキャッシュサイズ（MB）	4
データ転送速度（MB/s）	134
Active 時消費電力（Watt）	11.1
Idle 時消費電力（Watt）	7.5
Standby 時消費電力（Watt）	0.8
Spin-down 時消費エネルギー（Joule）	35.0
Spin-up 時消費エネルギー（Joule）	450.0
Spin-down 時間（sec）	0.7
Spin-up 時間（sec）	15.0

4. 実　　験

シミュレーション実験は，3. 節で説明したシミュレータを用いて RAPoSDA および MAID について電力削減率と性能に関して検証を行う．電力削減率については，省電力化手法を用いない同数のディスクドライブで構成されるストレージ（以下 Normal と記す）に対する消費電力量の比率で表すこととする．

4.1 シミュレーション構成

シミュレーションで用いるハードディスクドライブのモデルは Hitachi Global Storage Technologies の Hitachi Deskstar 7K2000 [7] に基づき，表 2 に示すパラメータを使用する．

シミュレーションは 12 時間分の人工ワークロードを用いて行う．各ストレージ構成におけるデータディスクは 128 台とし，キャッシュディスクは RAPoSDA と MAID の両手法でそれぞれ 1 台とする．人工ワークロードは読み出し比率を 70%，50%，30%と変化させた場合についてシミュレーションを実施する．

シミュレーションを行うストレージの構成は表 3 の通りであ

表 3　シミュレーションにおけるストレージ構成
Table 3　Configuration of storage system on this simulation

# of disks	Normal		MAID		RAPoSDA	
128	Cache Mem	-	Cache Mem	-	Cache Mem	16GB
	Cache Disks	-	Cache Disks	1	Cache Disks	1
	Data Disks	128	Data Disks	128	Data Disks	128

表 4　人工的ワークロードの諸元
Table 4　Parameters of synthetic workload

workload parameter	value
時間	約 12 時間
read:write	7:3, 5:5, 3:7
格納ファイル数	1,000,000 (1MB/file)
格納ファイルサイズ	2TB (Primary × Backup)
リクエスト数	$\lambda \times 3600 \times 12$
アクセス分布	Zipf 分布
アクセス到着分布	Poisson 到着
Zipf 係数 s	1.2
平均到着率	25 (request/sec)

る．表 3 において，ハイフンはそのストレージ構成では使用しないデバイスであり，Normal においてはキャッシュメモリおよびキャッシュディスクは使用せず，MAID ではキャッシュメモリを使用しないことを示している．

RAPoSDA におけるキャッシュメモリは，表 3 に示す容量のキャッシュメモリを全体の容量とし，電源系統が 2 系統の場合についてシミュレーションを行う．なおキャッシュメモリ上のデータは異なる電源系統のキャッシュメモリ間でプライマリデータとバックアップデータを分散して保持する．

4.2　人工ワークロードによる検証

シミュレーションで用いる人工ワークロードの諸元を表 4 に示す．生成するワークロードは Zipf 分布に基づくアクセスの偏りと，ポアソン過程に基づく到着間隔を持つ．ワークロードは read と write の比率をそれぞれ 7:3，5:5，3:7 の三つのパターンで生成し，各パターンについてシミュレーション実験を行う．

4.2.1　電力削減率の検証

シミュレーション実験における各ストレージ構成の消費電力量を図 5 に示す．グラフでは電力をエネルギーの単位（Joule）に変換して表しており，Normal の消費エネルギーはデータディスクの値，RAPoSDA および MAID の場合はデータディスクとキャッシュディスクの合計値としている．RAPoSDA と MAID の Normal に対する電力削減率は図 6 に示しており，両手法におけるデータディスクのスピンアップ回数は図 7 に示している．

図 5 の消費エネルギーをみると，RAPoSDA は読み出し比率に関係なく Normal に対して消費エネルギーを抑制している．MAID でもほぼ同じ状況であるが，読み出し比率が 70%の場合，経過時間が 2 時間前後では Normal よりも消費エネルギーが高くなっている．

図 6 より，読み出し比率が高い場合には RAPoSDA および

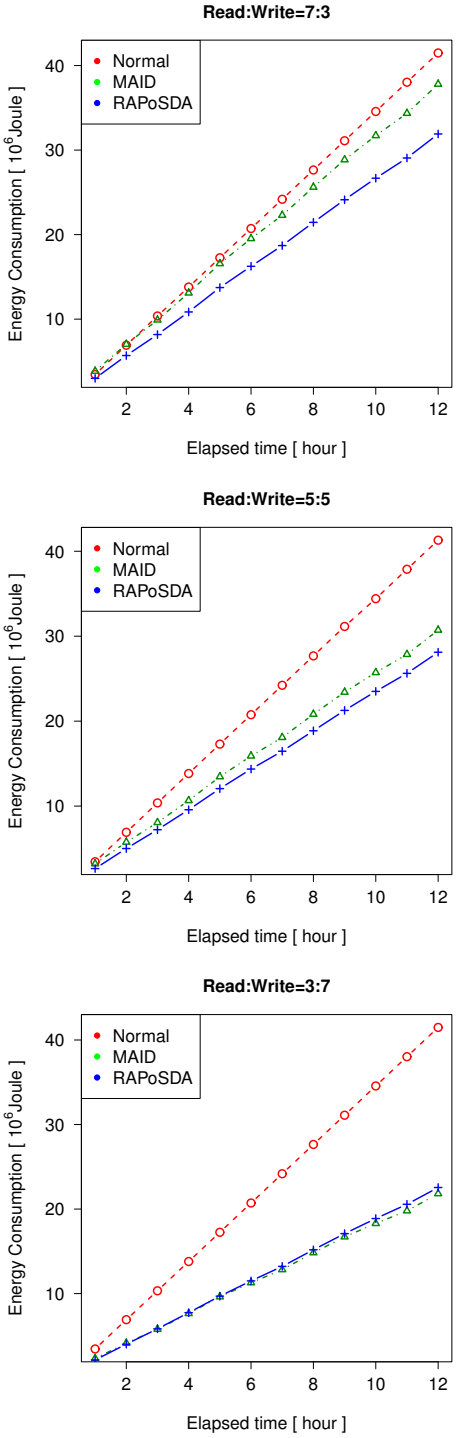

図 5　消費エネルギー
Fig. 5　Energy Consumption

MAID の両手法とも Normal に対する電力削減率の割合が小さくなることが分かる．これは読み出し比率が高いとキャッシュ（キャッシュメモリ or キャッシュディスク）における読み出しヒットミス率が高まり，Normal と同様にデータディスクへアクセスすることが多い為である．ただし運用時間が長くなるほどキャッシュの効果が得られて消費エネルギーは抑制されていく．

RAPoSDA と MAID の電力削減率を比較すると，読み出し比率が 50%以上においては，RAPoSDA の電力削減率は MAID よ

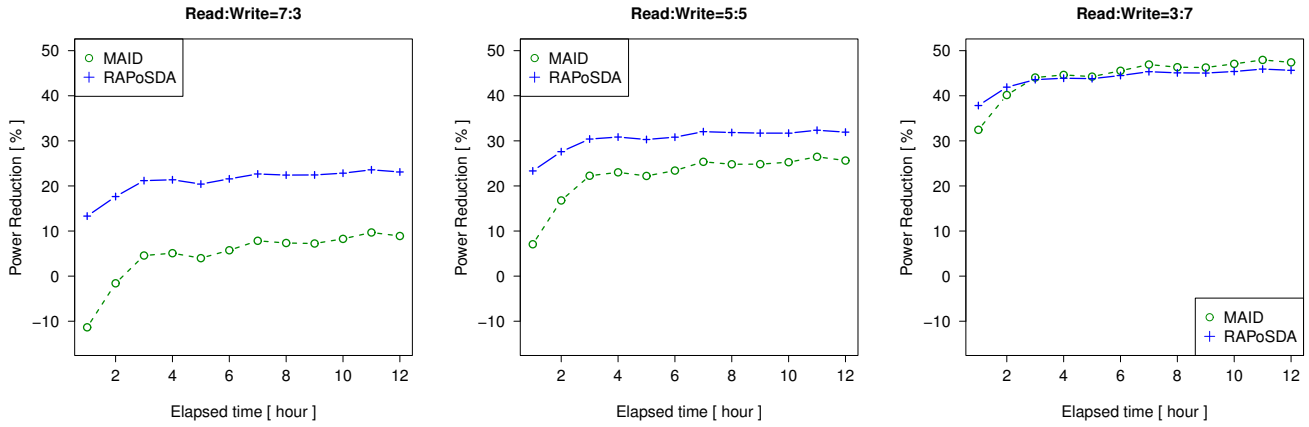

図 6　RAPoSDA および MAID における電力削減率
Fig. 6　Power Reduction of RAPoSDA and MAID

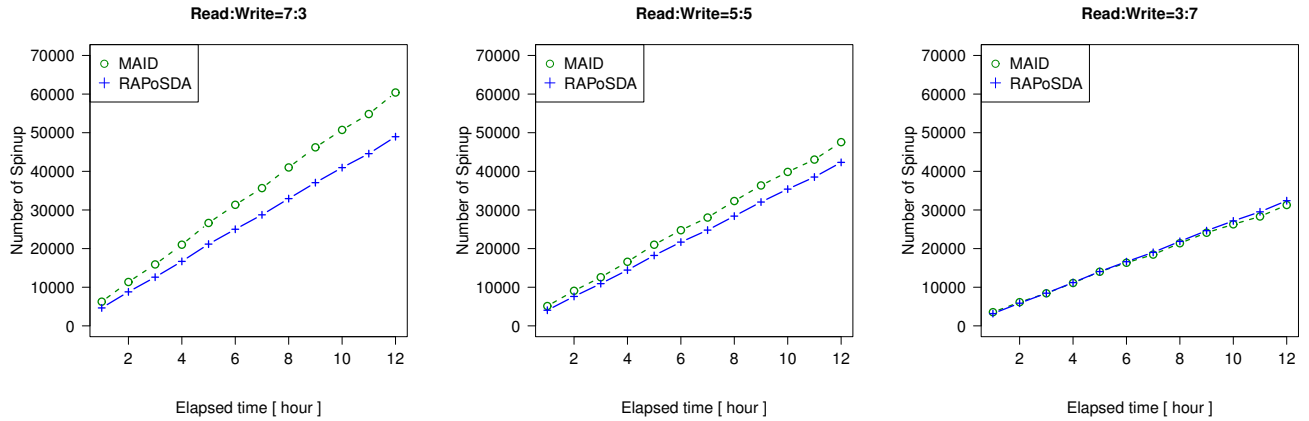

図 7　RAPoSDA および MAID におけるスピンアップ回数
Fig. 7　Number of Spinup in RAPoSDA and MAID

りも高く，より省電力効果得られている．しかし読み出し比率が低下するほど両手法の電力削減率の差は小さくなり，30%の読み出し比率においては長時間運用すると MAID の方が若干だが省電力効果が高くなっている．

これは RAPoSDA では書き込み比率が高いとキャッシュメモリのバッファ容量オーバーフローの発生頻度が多くなってしまうことが原因だと考えられる．図 8 は RAPoSDA において時間経過に伴ってキャッシュメモリのバッファ容量オーバーフローが発生した回数を表したグラフであるが，読み出し比率が 70%の時に比べ 30%では多くの時間帯でバッファ容量のオーバーフローが発生している．バッファ容量のオーバーフローが発生するとデータディスクへのアクセス頻度が高くなり全体のスピンアップ回数が高まってしまう恐れがある．図 7 の両手法のスピンアップ回数ををみると，やはり読み出し比率が低い（30%）場合では RAPoSDA のスピンアップ回数が MAID を上回っている．

この問題の対策としてはキャッシュメモリの容量を増やすことや，キャッシュメモリの系統を現在の 2 系統から複数系統に増やし，一つのキャッシュメモリに対応するデータディスク数を少なくしてオーバーフロー時におけるディスクアクセスの範囲を小さくすること等が考えられる．

電力削減率の全体傾向としては，実験開始から 4 時間前後までは削減率は上昇するが，それ以降では殆ど変化しなくなる．また図 5 と図 7 をみると消費エネルギーとスピンアップ回数には強い相関関係があることが確認できる．複数ディスクを用いるストレージにおいてはディスクをスタンバイ状態に移行させつつ，それらのディスクのスピンアップ回数をいかに抑制するかが重要なポイントとなる．

今回のシミュレーション実験では RAPoSDA，MAID ともにキャッシュヒット率は 90%以上であり，ディスクアクセスが発生した時点におけるデータディスクの回転確率は非常に高かった．我々が文献 [10] で構築した消費電力量の概算式では，キャッシュヒット率が高く（90%以上），ディスクアクセス時における回転中ディスクの割合が大きい（50%以上）場合では Normal に対する電力削減率は 20%から 40%になると見積もっており，シミュレーション実験での結果と良く一致していた．今回の実験結果から，ワークロード下における影響を考慮しても概算式による消費電力量の見積もりは有用であることが確認出来た．

4.2.2　性能の検証

本節では性能について検証する．図 9 は読み出し比率を変更

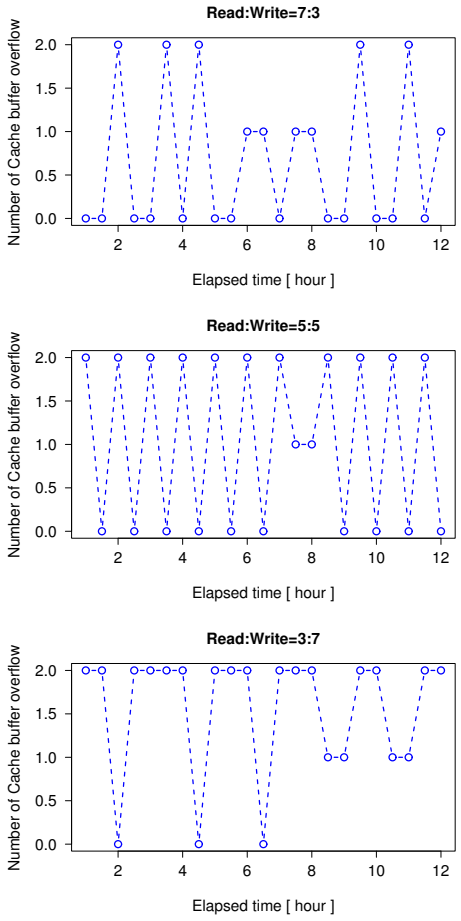

図 8　RAPoSDA におけるキャッシュメモリのバッファ容量オーバーフロー回数
Fig. 8　Number of cache memory buffer overflow in RAPoSDA

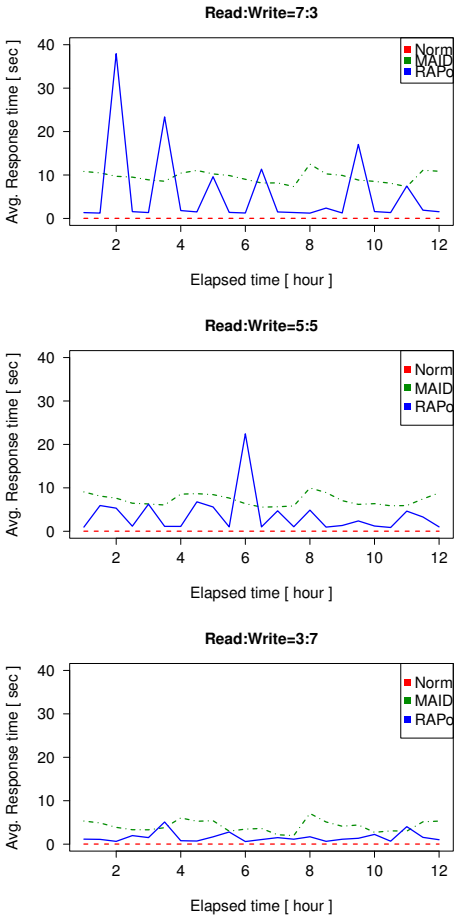

図 9　平均応答時間
Fig. 9　Average Response time

した場合における各ストレージ構成の平均応答時間である．平均応答時間は 30 分間隔でそのあいだに発生したリクエストの応答時間の平均で表している．

またシミュレーションの実行時間全体における平均応答時間は表 5 に示している．

図 9，図 5 をみると，Normal の平均応答時間が最も早いが，これはディスクのスピンアップやスピンダウンによる遅延が発生していないためである．応答性能は最も良いが，その反面消費電力量は最も高い．

読み出し比率が高い場合，RAPoSDA では時間帯によって応答時間にばらつきが生じている．これは図 8 におけるキャッシュメモリのバッファ容量オーバーフローのタイミングと相関がある．キャッシュメモリのバッファがあふれた時にはデータディスクへの書き込みが発生するが，このときバッファに蓄積されていたデータに対応する全てのデータディスクへアクセスを試みるため，この時間帯における平均応答時間が劣化してしまう．経過時間が少ない時ほど応答時間の劣化が大きいが，これはキャッシュの読み出しヒットミスによるディスクアクセスとの複合的な影響によるものと考えられる．

一方読み出し比率が少なくなると平均応答時間は早くなっている（図 9，Read:Write=3:7）．図 8 をみると読み出し比率が 30%のときは多くの時間帯でキャッシュメモリのバッファ容量オーバーフローが発生しており，そのためデータディスクへのアクセス頻度が高まりディスクが回転している間にアクセスしている割合が高まることが影響していると考えられる．

MAID では，RAPoSDA ほど応答時間の変動は少ないが全体の平均応答時間は最も遅い．読み出し比率が少なくなるほど平均応答時間は早くはなるが RAPoSDA の方が全体の平均応答時間は早いことが確認できる．

表 5　各ストレージ構成における平均応答時間 [秒]
Table 5　Average Response time of each configuration [sec]

Read:Write	Configuration		
	Normal	MAID	RAPoSDA
7:3	0.02	9.55	5.75
5:5	0.02	7.25	3.70
3:7	0.02	4.14	1.57

5. 関 連 研 究

PARAID [8] は RAID 構成のストレージシステムに対する省電力化手法であり，データのストライピングのパターンを偏らせてアクセスの無いディスクを作り，そのディスクを停止させ

る．従来の RAID に対して性能劣化はほとんどなくある程度の省電力効果は得られるが，PARAID も RAID 構成に特化しているため，RAID 構成でないストレージに対しては適用できない．

GRAID [6] も RAID 構成ストレージの省電力化手法の一つである．GRAID では省電力化と信頼性の確保に重点をおいており，RAID10 ベースのディスクアレイを前提としている．通常のディスクドライブの他に，ログディスクというログ格納用のディスクを用いることにより，ミラーリングされたディスクペアの一方のディスクアクセスを抑制する．

EERAID [5] は RAID コントローラーレベルでの動的な I/O スケジューリングとキャッシュ管理ポリシーによって RAID 構成ストレージの省電力化を実現する手法である．RAID 構成に特化している．

ディスクドライブ単体を対象とした DRPM [3] では，ディスクドライブの消費電力量はその回転数（RPM）の関数として表せることを示し，負荷に応じて動的にディスクの回転数を変更することにより，省電力化と性能の維持を実現することを提案した．しかし回転数の動的な変更には技術的な課題も多く，多段階に渡って動的に回転数を変えることができるディスクドライブは未だに実用化はされていない．

6. まとめおよび今後の課題

今回我々は新規に構築したシミュレータを用いて，ストレージの省電力化手法である RAPoSDA および MAID についてその省電力効果と性能に関して検証を行った．人工的なワークロード下における検証を行った結果，消費電力，応答性能の両面で MAID に対し RAPoSDA の優位性を示すことができた．今後はデータディスク数やキャッシュディスク数を増やした場合での省電力効果や性能に関する検証を行い，さらに実環境での運用を想定して，外部公開されているファイルサーバー等のトレースをベースにしたワークロードによるシミュレーション実験を行う予定である．

また実際のハードディスクドライブにおける消費電力を計測し，シミュレータとの誤差を把握し，より精度の高いシミュレーション実験を行えるようにシミュレータを改善することも必要である．

その他としては信頼性に関して定量的な評価を行うことも今後の課題である．

謝　　辞

本研究の一部は，文部科学省科学研究費補助金特定領域研究 (#21013017)，日本学術振興会科学研究費補助金基盤研究 (A)(#22240005) の助成により行われた．

文　　献

[1] Dennis Colarelli and Dirk Grunwald. Massive arrays of idle disks for storage archives. In *Supercomputing '02: Proceedings of the 2002 ACM/IEEE conference on Supercomputing*, pp. 1–11, Los Alamitos, CA, USA, 2002. IEEE Computer Society Press.

[2] Greg Ganger, et. al. The disksim simulation environment (v4.0). http://www.pdl.cmu.edu/DiskSim/.

[3] Sudhanva Gurumurthi, Anand Sivasubramaniam, Mahmut Kandemir, and Hubertus Franke. Drpm: Dynamic speed control for power management in server class disks. *Computer Architecture, International Symposium on*, Vol. 0, p. 169, 2003.

[4] Hui-I Hsiao and David J. DeWitt. Chained declustering: A new availability strategy for multiprocessor database machines. In *Proceedings of the Sixth International Conference on Data Engineering*, pp. 456–465, Washington, DC, USA, 1990. IEEE Computer Society.

[5] Dong Li and Jun Wang. Eeraid: energy efficient redundant and inexpensive disk array. In *EW 11: Proceedings of the 11th workshop on ACM SIGOPS European workshop*, p. 29, New York, NY, USA, 2004. ACM.

[6] Bo Mao, Dan Feng, Suzhen Wu, Lingfang Zeng, Jianxi Chen, and Hong Jiang. Graid: A green raid storage architecture with improved energy efficiency and reliability. In *MASCOTS*, pp. 113–120, 2008.

[7] Hitachi Global Storage Technologies. Hard disk drive specification, hitachi deskstar 7k2000. http://http://www.hitachigst.com/tech/techlib.nsf/techdocs/5F2DC3B35EA0311386257634000284AD/$file/USA7K2000_DS7K2000_OEMSpec_r1.4.pdf.

[8] Charles Weddle, Mathew Oldham, Jin Qian, An-I Andy Wang, Peter Reiher, and Geoff Kuenning. Paraid: A gear-shifting power-aware raid. *Trans. Storage*, Vol. 3, No. 3, p. 13, 2007.

[9] John Zedlewski, Sumeet Sobti, Nitin Garg, Fengzhou Zheng, Arvind Krishnamurthy, and Randolph Wang. Modeling hard-disk power consumption. In *Proceedings of the 2nd USENIX Conference on File and Storage Technologies*, pp. 217–230, Berkeley, CA, USA, 2003. USENIX Association.

[10] 引田論之, 横田治夫. プライマリ・バックアップ構成を有効利用したストレージシステムの省電力化手法の提案. In *DEIM Forum 2010 E6-4*, 2010.