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

引田　諭之† 　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
†† 東京工業大学学術国際情報センター　〒 152–8550 東京都目黒区大岡山 2–12–1
E-mail: †hikida@de.cs.titech.ac.jp, ††, †yokota@cs.titech.ac.jp

あらまし　近年における情報量の爆発的な増加に伴い，ストレージシステムにおける消費電力量の増加が大きな問題となっている．本研究では，ストレージシステムにおいて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りデータ配置を工夫することによって，高信頼で優れた省電力効果を実現する手法を提案する．手法の有効性を検証するために消費電力量を概算する理論式を構築し，その効果の見積もりを行う．
キーワード　省電力，ストレージ，ディスクドライブ

An Energy Reduction Method for Disk Storage Systems Utilizing the Primary-Backup Configuration

Satoshi HIKIDA† 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
†† Global Scientific Information and Computing Center, Tokyo Institute of Technology
2–12–1 Ookayama, Meguro-ku, Tokyo 152–8550, Japan
E-mail: †hikida@de.cs.titech.ac.jp, ††, †yokota@cs.titech.ac.jp

Abstract　The power consumption of disk storage systems in a data center is now a big problem corresponding to the information explosion in recent years. In this paper, we propose a method for reducing the power consumption of the storage systems with keeping the reliability of data by utilizing the primary-backup configuration. To verify the effectiveness of the proposed method, we developed formulas calculating power consumption of the storage system roughly. The estimation results with varying the parameters indicate that the proposed method is effective.
Key words　Energy saving, Storage, Disk drive

1. はじめに

近年のデータセンターでは消費電力量が重要な問題となっている．杉浦 [6] の報告によると，2000 年から 2006 年にかけての全米におけるデータセンターの消費電力量は約 28 億 kWh から約 61 億 kWh と，6 年間でおよそ 2 倍となっており，年率 19%で増加している．また，2006 年における消費電力量の内訳では，IT 機器（サーバー，ネットワーク機器，ストレージ）の占める割合は 47%であり，今後も増加が見込まれる．このような状況から，IT 機器に対する省電力化は重大な課題となっている．

IT 機器のうちストレージの消費電力量も増加しており，2000 年から 2006 年にかけては約 3 倍の増加となっている [6]．さらに今後は SNS(Social Networking Services) や SaaS(Software as a Service)，クラウドコンピューティングなどのインターネット技術を利用した新しい情報基盤の登場や，昨今の企業のコンプライアンス上の問題などにより，記録すべきデータ，保持すべきデータが急増してきており，ストレージに対する省電力化は重要な課題である．

このような状況から，従来よりストレージの省電力化については様々な研究がなされてきた．しかし，それらは主に省電力化と性能の維持にのみ着目したものが多く，ストレージの重要な要素である信頼性についてはあまり考慮していなかった．そこで，本研究ではストレージの信頼性の確保にも着目した新しい省電力化手法を提案する．

この提案手法ではキャッシュメモリとディスクドライブの双

方でプライマリデータとバックアップデータを保持する構成をとる．そしてデータ配置やディスクアクセスの制御などを工夫することによりストレージの省電力化を実現する．

本論文の構成は以下の通りである．2. ではディスクドライブの動作のモデルと消費電力の概要を説明する．3. では提案手法の構成について詳細な説明を行い，4. では消費電力量を見積もるための概算式を示し，パラメータを変更することにより様々な状況での消費電力削減率を見積もる．5. では関連研究について述べ，6. でまとめと今後の課題について述べる．

2. ディスクドライブの動作モデル

ディスクドライブは様々な部品から構成されており，部品によって消費される電力は異なる．以下ではディスクドライブを構成する部品のうち，消費電力に影響を与えるものに着目し，ディスクドライブの動作を状態遷移の観点から述べて，ディスクドライブの振る舞いとその消費電力について述べる．

2.1 ディスクドライブの構成

ディスクドライブを構成する主要な部品は，データを記録するディスク（またはプラッター）と呼ばれる円盤と，データの読み取り/ 書き込みを行うヘッドを搭載したアーム，モーター駆動によりプラッターを回転させるスピンドル等がある．ディスクドライブはこれらの機械部品と，それらの動作を制御するコントローラー等の電子部品とから構成されている．

2.2 ディスクドライブの状態遷移

ディスクドライブには大きく分けて以下の３つの状態がある [3] [5].

- アクティブ（Active）状態

ディスクドライブが入出力処理を行っている状態であり，スピンドルモーターは最高回転数で回転しており，ヘッドはディスク上に存在し，データ転送を行っている． 消費電力量（P_{read} および P_{write}）は他の状態に比べ一番大きい．

- アイドル（Idle）状態

ディスクドライブは入出力処理は行っていないが，スピンドルモーターは最高回転数で回転中であり，入出力要求に対して迅速に対応出来る状態である．ヘッドはディスク上に存在している．消費電力量（P_{idle}）はアクティブ状態よりは低いが，スタンバイ状態よりも大きい．

- スタンバイ（Standby）状態

ディスクドライブは入出力処理を行っておらず，スピンドルモーターも停止しており，ヘッドはランプと呼ばれるプラッター外の機構に退避している状態である．消費電力（$P_{standby}$）は３状態の中で一番小さいが，入出力要求の発生時には一時的にだがアクティブ状態よりも大きな電力（P_{tran}）を消費する．また，スピンドルモーターが最高回転数で回転するまでの時間およびヘッドがランプからディスク上に移動した後にデータをシークするため，時間的な損失も大きい．

ディスクドライブは図 1 に示すように，各状態間を遷移する．

2.3 ディスクドライブの消費エネルギー

ディスクドライブの消費エネルギーは，スピンドルモーターと装置電源がほとんどを占めている [1]．特に大きいのがスピンドルモーターで，ディスクの回転を停止させるとき（アクティブ状態からスタンバイ状態への遷移時）と，回転を開始させるとき（スタンバイ状態からアイドル状態への遷移時）に，非常に大きいエネルギー（アクティブ状態時よりも大きなエネルギー）を消費する．

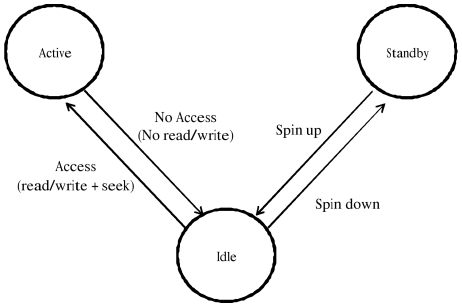

図 1　ディスクドライブの状態遷移図

各状態および状態遷移で消費電力量を比べてみると，一番大きいのはディスクが回転を開始する時，すなわちスタンバイ状態→アイドル状態への遷移時である．次に大きいのはアクティブ状態で，さらにアイドル状態，スタンバイ状態の順となる．

ディスクドライブの振る舞いから考えると，消費エネルギーを削減するためには長期間に渡るスタンバイ状態をつくることが重要である．ただし，ディスクの回転を停止させる為に無闇にスピンダウンを行うと，もしディスクアクセスが発生した場合はディスクのスピンアップが必要となり，スピンアップに伴うエネルギー損失が非常に大きいため，消費電力量はかえって従来よりも大きくなってしまう恐れもある．

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．

3. 提案手法について

3.1 構　　成

2.3 で述べた要求を満たすための手法として，本論文ではキャッシュメモリとディスクドライブでプライマリ・バックアップ構成をとり，データ配置を工夫することで省電力化と信頼性の確保を実現する手法を提案する．

本提案手法は以下の要素から構成される（図 2）.

- キャッシュメモリ
- データディスク
- キャッシュディスク

3.1.1 キャッシュメモリ

本提案手法ではキャッシュメモリは揮発性の記憶媒体であることを前提としているため，ノードの故障が発生した場合にはキャッシュメモリ上のデータが消失してしまう恐れがある．こ

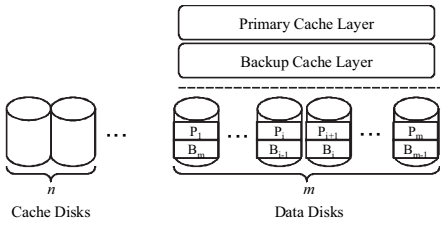

図 2　提案手法の全体構成

の問題に対応するためにキャッシュメモリ上のデータに冗長性を持たせる．

具体的には，キャッシュ単位毎にプライマリのデータを保持するプライマリ層と，バックアップデータを保持するバックアップ層を構成する．あるキャッシュ単位のプライマリ層に書き込まれたデータは，別のキャッシュ単位のバックアップ層にバックアップデータを書き込む．ただし，それぞれのキャッシュ単位は物理的に異なるノードで構成され，各ノードで個別の電源を有し UPS 等の断電対策が施されているものとする．このプライマリ・バックアップ構成は Chained Declustering [4] という手法に基づいたデータ配置方法である．

また，キャッシュメモリの効果はキャッシュサイズに依存すると考えられる．そのため，本提案手法では複数のノードでメモリを共有する構成を考える．メモリを共有するノード数を一つの単位としてデータの管理方法を考えるため，この単位のことをキャッシュ単位と呼び，キャッシュ単位を構成するノード数を N_{cu} で表す．

$N_{cu} = 1$ では，メモリとノードは 1 対 1 で対応しているので，これを単純構成（図 3）と呼び，$N_{cu} \geqq 2$ では複数のノードでメモリを共有するのでこれを複数構成（図 4）と呼ぶ．

複数構成では，キャッシュメモリを複数ノードで共有するため，単純構成よりも大きい容量のメモリバッファを柔軟に利用できる．例えば，アクセス頻度が極端に多いノードと，アクセス頻度が極端に少ないノードでキャッシュ単位を構成すれば，ほとんどのデータはアクセス頻度の多いディスクのものなので，２ノード分のメモリバッファを実質 1 台のノードで使用しているように振る舞う．メモリバッファの効率はその容量に大きく依存するため，このような組み合わせでは優れたキャッシュ効果が期待できる．

3.1.2　データディスク

実際のデータを保持するディスクドライブ群である．ディスクアクセスが，ある一定の閾値時間を超えて発生しなかった場合，そのディスクドライブの回転を停止させ，スタンバイ状態とする．スタンバイ状態を長く維持し続けられれば，それだけ消費電力の削減に寄与する．また，ディスクドライブでも信頼性の確保のためにプライマリ・バックアップ構成をとる．

3.1.3　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ドライブ群である．キャッシュメモリでは大規模データを扱うディスクストレージ群に対してはキャッシュ効果が限られてしまう．特に大量の読み込み要求が発生する負荷においてはディスクアクセス頻度を抑制する効果があまり期待出来ない．そのため，キャッシュメモリよりも容量が格段に多い少数のディスクドライブを読み出し処理専用のキャッシュとして用いることにより，データディスクへのアクセス頻度を低く抑える．キャッシュディスクを用いる手法は MAID [1] で提案されている．

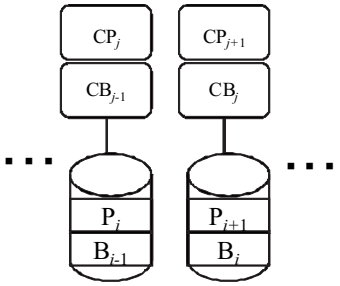

図 3　構成例 $N_{cu} = 1$

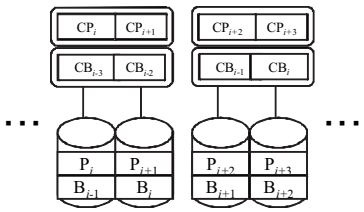

図 4　構成例 $N_{cu} = 2$

3.2　動　　作

3.2.1　書き込み処理

書き込み時は，キャッシュメモリ，データディスク，キャッシュディスクの順に処理を行う．

書き込みデータを d_w とする．d_w を格納しているディスクが所属しているキャッシュ単位のプライマリ層に d_w を書き込む．それと同時に別のキャッシュ単位を構成しているバックアップ層に d_w のコピーデータを書き込む．

もしバッファが閾値を超えた場合は，データディスクに実際に書き込み処理を行う．このとき，バッファ中で一番多くデータをキャッシュしているデータディスクを処理対象とする．該当データディスクが回転中の場合，そのままキャッシュデータを書き込む．もし該当データディスクが停止中であれば，データディスクをスピンアップし，その後でキャッシュデータを書き込む．データディスクに書き込む際，もし書き込みデータがディスクのプライマリデータであれば，同一ディスクに格納するべきバックアップデータをキャッシュしているキャッシュ単位からそのデータも同時にデータディスクに書き込む．こうすることでプライマリデータの書き込みとバックアップデータの書き込みが異なるタイミングで発生することを防ぎ，無駄なディスクアクセスを低減させてディスクのスピンアップに伴う大きなエネルギー消費を抑制することができる．

データディスクに書き込んだデータはそのままキャッシュディスクにコピーする．その理由は，次回同じデータに対する読み出し要求が発生した場合に，データディスクへのアクセスを抑制させるためである．

3.2.2 読み出し処理

キャッシュメモリは主に書き込みデータに対するバッファとして用いているため，読み出し要求に対するプリフェッチとしての働きは少ない．そのため，書き込みバッファとして用いていると読み出し時のキャッシュヒット率は著しく低下してしまい，該当データディスクへのアクセスが多くなってしまう．データディスクへのアクセスが多くなるとデータディスクが回転を停止させるための閾値時間の確保が困難になってしまったり，データディスクのスピンアップが頻繁に発生してしまい，省電力効果が期待できず，最悪の場合は通常の運用時よりも大きなエネルギーを消費してしまう可能性がある．

そこで，本提案手法では MAID [1] と同様のアプローチをとり，読み出し要求に対するキャッシュとして少数のディスクを用意する．これらのディスクはキャッシュディスクと呼び，常にアクティブ状態を維持し続け回転が停止することはない．しかし，本提案手法ではキャッシュディスクにデータを配置するステージング方法が MAID とは異なる．

読み出し要求に対する処理としては，まずキャッシュメモリに該当データがあるかを確認し，あればキャッシュメモリから読み出す．もしキャッシュメモリに存在しなければ次にキャッシュディスクを確認し，キャッシュディスク中に該当データが存在すればそこから読み出す．

キャッシュディスク中にデータが存在しない場合は，データディスクにアクセスする．該当データはプライマリディスクとバックアップディスクのどちらかから読み出せばよいので，各ディスクが回転中かどうかをまずは確認する．どちらか一方のみが回転中の場合，そちらのディスクからデータを読み出す．両方とも回転中の場合，メモリバッファのキューが長い方のディスクから読み出す．

両方とも停止中だった場合，まず両ディスクの回転停止時間を調べ，ブレイクイーブン時間を超えているかどうかを確認する．もし一方のディスクのみが超えていた場合，そちらのディスクをスピンアップさせる．両ディスクともにブレイクイーブン時間を超えて停止していた場合，それぞれの回転停止時間の長さを確認する．回転停止時間が異なる場合は，停止時間の長かった方のディスクをスピンアップさせる．停止時間が両ディスクで同じだった場合，メモリバッファのキューが長い方のディスクを調べ，そちらのディスクをスピンアップさせる．

データを読み出した後は，対応するメモリバッファのデータをディスクに書き込む（プライマリ層，バックアップ層の両方から）．この時点でメモリバッファ上のデータを書き込むことによってメモリバッファの容量に余裕が出来るため，次の書き込み要求時にディスクアクセスが発生する頻度を低減させることが可能となる．さらに，異なるキャッシュ単位のキャッシュメモリ上に格納されているプライマリ層とバックアップ層のデータを同期して書き込むことで，プライマリ層とバックアップ層のデータがそれぞれ異なるタイミングでメモリバッファの閾値を超えた場合に比べ，ディスクのスピンアップ頻度を低減させることができる．

その後，読み出し要求に対するキャッシュヒット率を向上させるために，読み出したデータおよびデータディスクに書き込んだデータをキャッシュディスクに書き込む．

3.3 提案手法の拡張

回転数（RPM）を負荷に応じて動的に変更出来るディスクが研究されている [3]．これはディスクアクセスが無く，アクティブ状態からスタンバイ状態に遷移する際，ディスクの回転を停止させる変わりに低速な回転数で回転させることにより省電力化を実現する技術である．回転を停止させないため，従来のディスクに比べたらスタンバイ時の消費電力（$P_{standby}$）は大きくなるが，ディスクアクセスが発生した時にスピンアップさせることによって生じる高い消費電力（P_{tran}）を大幅に削減でき，場合によっては回転を停止させる方法よりも高い省電力効果をもたらす．

さらに最高回転数に到達するまでの時間も従来のディスクよりも短いため，応答時間の損失も少なく済むという利点がある．本提案手法でもこのような低速回転ディスクを採用することは容易である．

また，ストレージの信頼性をより高めるために，データの冗長性を多重化する構成も考えられる．キャッシュ単位毎にプライマリデータとバックアップデータをもつ場合は二重化構成だが，このバックアップが他の 1 つのキャッシュ単位のものではなく，他の複数のキャッシュ単位に対するバックアップを持つことにより，三重化やそれ以上の多重化も構成でき，より信頼性の高いストレージを実現できる．

4. 消費電力量の概算

本提案手法では，3. で述べたようにキャッシュメモリ，データディスク，キャッシュディスクで構成されている．消費電力量の概算式は，データディスクでの消費電力，キャッシュディスクでの消費電力をそれぞれ構築し，その合計が本提案手法における総消費電力量とする．

以下ではデータディスク，キャッシュディスクにおける消費電力量の概算式を示し，その次に本提案手法を用いない従来方式における消費電力量の概算式を示す．概算式で用いる記号とその意味は表 1 に示す．

4.1 提案手法における概算式

4.1.1 データディスク

まず書き込み処理時について考える．書き込み要求頻度を f_w，バッファ書き込み可能率を b_w とすると，データディスクの総数 m に対して $mb_w f_w$ のディスクはメモリバッファまでのアクセスのためディスクアクセスは発生しない．このとき，r_d の割合でディスクが回転中だとすると，回転中のディスクの割合は mr_d で，停止は $(1 - r_d)m$ となる．このときのデータディスクの消費電力量は $mb_w f_w(r_d P_{idle} + (1 - r_d)P_{standby})$ となる．また，メモリバッファのキューが閾値を超えていた場合，データディスクに直接書き込むため，その時の消費電力量は $mf_w(1 - b_w)(r_d P_{write} + (1 - r_d)(P_{write} + P_{tran}))$ となる．

次に読み込み処理時について考える．読み込み要求頻度を f_r，メモリキャッシュのヒット率を h_c，キャッシュディスクのヒット率を h_d とすると，データディスクにアクセスが無い時の消費

表 1　概算式で用いる記号とその説明

記号	説明
n	キャッシュディスク数
m	データディスク数
$P_{standby}$	スタンバイ状態時のディスク消費電力
P_{idle}	アイドル状態時のディスク消費電力
P_{tran}	スタンバイ状態からアイドル状態へ遷移する際の消費電力
P_{read}	アクティブ状態（read）時のディスク消費電力
P_{write}	アクティブ状態（write）時のディスク消費電力
$P_{dataDisk}$	データディスク全体の消費電力
$P_{cacheDisk}$	キャッシュディスク全体の消費電力
P_{normal}	従来方式のストレージ全体の消費電力
h_c	読み出しアクセスに対するキャッシュメモリのヒット率
h_d	読み出しアクセスに対するキャッシュディスクのヒット率
b_w	メモリバッファに対する書き込み可能率
r_d	ディスクアクセス時にディスクが回転している確率

表 2　ディスクドライブの各状態における消費電力量

記号	値 [Watt]
P_{read}	13
P_{write}	13
P_{idle}	9.3
$P_{standby}$	0.8
P_{tran}	24

電力は $m(1-h_c)h_d f_r(r_d P_{idle}+(1-r_d)P_{standby})$ となる．また，メモリ，キャッシュディスクの両方でキャッシュヒットミスした場合，直接データディスクから読み出すため，その時の消費電力は $f_r m(1-h_c)(1-h_d)(r_d P_{read}+(1-r_d)(P_{read}+P_{tran}))$ となる．

従って，データディスクの消費電力の合計は以下の式で表せる．

$$\begin{aligned} P_{dataDisk} = & \\ & m(h_c f_r+(1-h_c)h_d f_r+b_w f_w+(1-(f_r+f_w)))(r_d P_{idle}+(1-r_d)P_{standby}) \\ & + f_r m(1-h_c)(1-h_d)(r_d P_{read}+(1-r_d)(P_{read}+P_{tran})) \\ & + f_w m(1-b_w)(r_d P_{write}+(1-r_d)(P_{write}+P_{tran})) \end{aligned} \quad (1)$$

4.1.2　キャッシュディスク

キャッシュディスクでは基本的に読み込み処理時において電力消費が発生する．キャッシュディスクの総数を n とすると，キャッシュメモリにヒットした場合および読み出し処理自体が発生しなかった頻度 $(1-f_r)$ のときはアイドル状態のため，その消費電力は $n(h_c f_r+(1-f_r))P_{idle}$ となる．

また，キャッシュメモリがヒットミスした場合において，キャッシュディスクではヒットした時の消費電力は $f_r n(1-h_c)h_d P_{read}$ となり，キャッシュディスクでもヒットしなかった場合はデータディスクからデータを読み出し，さらにそのデータをキャッシュディスクに書き込むため $f_r n(1-h_c)(1-h_d)P_{write}$ の消費電力となる．

さらに，メモリバッファをオーバーフローしてデータディスクにデータを書き込む際も，該当データはキャッシュディスクに書き込まれるため $f_w n(1-b_w)P_{write}$ の消費電力も発生する．

これらより，キャッシュディスクの消費電力の合計は以下の式となる．

$$\begin{aligned} P_{cacheDisk} = & n(h_c f_r+(1-f_r))P_{idle} \\ & + f_r n(1-h_c)h_d P_{read} \\ & + f_r n(1-h_c)(1-h_d)P_{write} \\ & + f_w n(1-b_w)P_{write} \end{aligned} \quad (2)$$

4.1.3　提案手法における総消費電力

本提案手法における総消費電力量 P_{total} は以下のように概算できる．

$$P_{total} = P_{dataDisk} + P_{cacheDisk} \quad (3)$$

4.1.4　従 来 方 式

提案手法を用いない従来の方式では，キャッシュディスクは用いないため消費電力の概算式は次のようになる．

$$P_{normal} = f_r m P_{read} + f_w m P_{write} + m(1-(f_r+f_w))P_{idle} \quad (4)$$

4.2　MAID における概算式

提案手法との消費電力量削減率を比較するために，MAID に関しても同様に消費電力量の概算式を構築する．

4.2.1　データディスク（MAID）

MAID では，書き込み/読み出し処理の両方でキャッシュディスクを用いているため，キャッシュヒット率は h_d のみを用いる．データディスクに対してアクセスが発生しなかった場合の消費電力量は $m(h_d f_r+h_d f_w+(1-(f_r+f_w)))(r_d P_{idle}+(1-r_d)P_{standby})$ であり，データディスクに対して読み出しアクセスが発生している場合は $m f_r(1-h_d)(r_d P_{read}+(1-r_d)(P_{read}+P_{tran}))$ となり，書き込みアクセスが発生している場合は $m f_w(1-h_d)(r_d P_{write}+(1-r_d)(P_{write}+P_{tran}))$ となる．したがって，MAID におけるデータディスクの消費電力合計は以下の式となる．

$$\begin{aligned} P_{dataDisk(MAID)} = & \\ & m(h_d f_r + h_d f_w + (1-(f_r+f_w)))(r_d P_{idle} + (1-r_d)P_{standby}) \\ & + m f_r(1-h_d)(r_d P_{read} + (1-r_d)(P_{read} + P_{tran})) \\ & + m f_w(1-h_d)(r_d P_{write} + (1-r_d)(P_{write} + P_{tran})) \end{aligned} \quad (5)$$

4.2.2　キャッシュディスク（MAID）

MAID におけるキャッシュディスクは，読み出し/書き込みの両方の要求からアクセスされるため，ディスクがアイドル状態となるのはアクセスが発生しないときである．そのときの消費電力量は $n(1-(f_r+f_w))P_{idle}$ であり，ディスクに対するアクセスが発生した場合を考慮すると，以下の式となる．

$$P_{cacheDisk(MAID)} = n(1-(f_r+f_w))P_{idle} + n f_r h_d P_{read} + n f_w P_{write} \quad (6)$$

4.2.3　MAID における総消費電力

上記式（5），式（6）より，MAID における総消費電力量 $P_{total(MAID)}$ は以下のように概算できる．

$$P_{total(MAID)} = P_{dataDisk(MAID)} + P_{cacheDisk(MAID)} \quad (7)$$

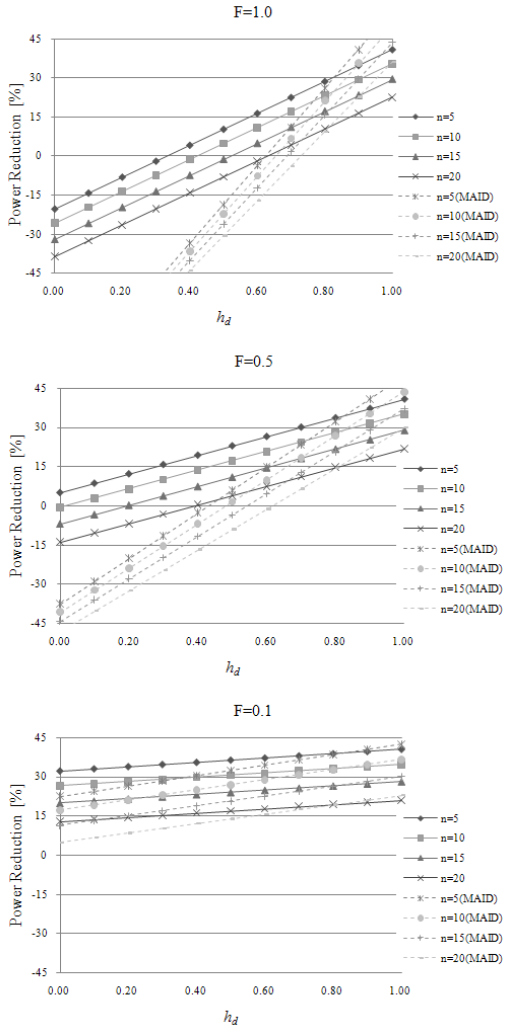

図 5　ディスクキャッシュのヒット率 h_d を変化させた場合での消費電力削減率の概算．次のパラメータは固定値としている（h_c= 0.5, b_w= 0.5, r_d= 0.5, f_r= 0.8, f_w= 0.2）

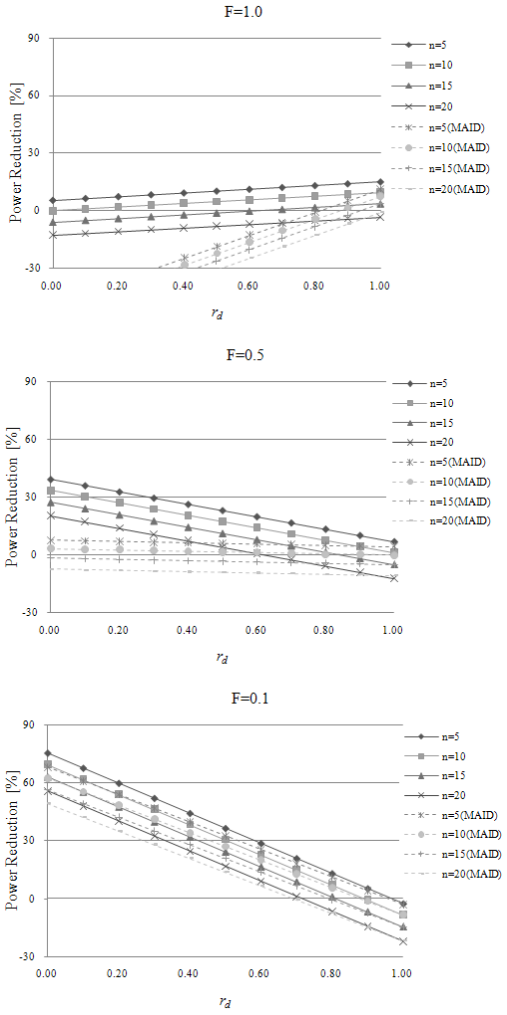

図 6　ディスクアクセス時での回転確率 r_d を変化させた場合での消費電力削減率の概算．次のパラメータは固定値としている（h_c= 0.5, h_d= 0.5, b_w= 0.5, f_r= 0.8, f_w= 0.2）

4.3　算出式を用いた見積もり

図 5，6 では 1 つのパラメータに着目し，そのパラメータを変更したときに提案手法と MAID のそれぞれがディスクストレージの消費電力量に与える影響をグラフに表している．

図 7，8 で示しているグラフでは，MAID の概算式では用いていないパラメータについて，同様にそのパラメータを変数として消費電力量の見積もりを行った結果を示している．グラフにおける消費電力削減率は従来方式における消費電力量に対する本提案手法での削減率を表している．グラフ中，高負荷とはアクセス頻度の高い場合を示しており，これはパラメータの f_w と f_r の合計 F（$F = f_r + f_w$）が 1 として計算している．中負荷は同様に F = 0.5 で，低負荷は F = 0.1 として算出している．

また，ディスクの総数を 100 台と固定し，そのうちキャッシュディスク n の台数を 5 台から 20 台まで変更した場合における消費電力削減率を求めた（データディスク m の台数は 95 台から 80 台）．

概算に用いる各状態や状態遷移時における消費電力量の値は表 2 に示している．これは Seagate ST3500630AS のディスクドライブのモデルのもので，他の研究 [2] でも用いられている．

4.4　消費電力量の概算における考察

図 5 をみると，キャッシュディスクのヒット率 h_d が高くなると MAID の省電力効果は提案手法より若干高くなるが，これは r_d を固定して計算しているためで，実際は h_d が高くなるとディスクの回転確率 r_d は小さくなるため，実運用環境では省電力効果は提案手法の方がより効果的であると予想される．しかし，キャッシュヒット率が高くない場合では，提案手法の方が優れた省電力効果があるといえる．また，図 6 では負荷の程度に関わらず提案手法が MAID よりも省電力効果は高いことが示されている．実際には，r_d が上がると h_d が下がるので，提案手法の優位性がさらに増すと考えられる．

図 5，図 7，図 8 はデータのキャッシュ率を表しているため，キャッシュ率が高いほど省電力効果は当然高くなる．高負荷時においてはその効果が顕著にあらわれている．また，すべてのグラフにおいて，キャッシュディスクの台数は少ない方が省電力効果は高いことが分かる．現実には総ディスク数に対するキャッシュディスク数の割合はキャッシュヒット率とのトレード

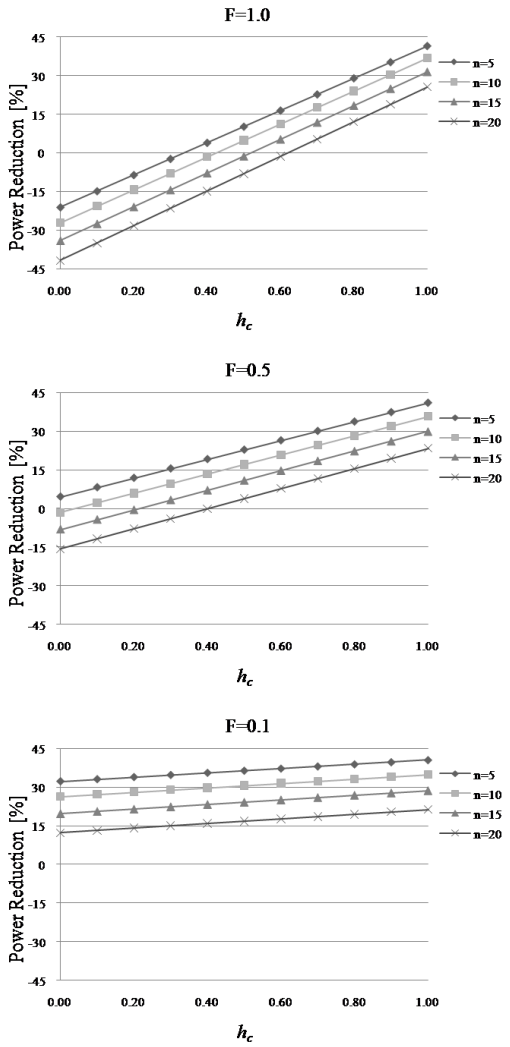

図 7　キャッシュメモリのヒット率 h_c を変化させた場合での消費電力削減率の概算．次のパラメータは固定値としている（h_d= 0.5, b_w= 0.5, r_d= 0.5, f_r= 0.8, f_w= 0.2）

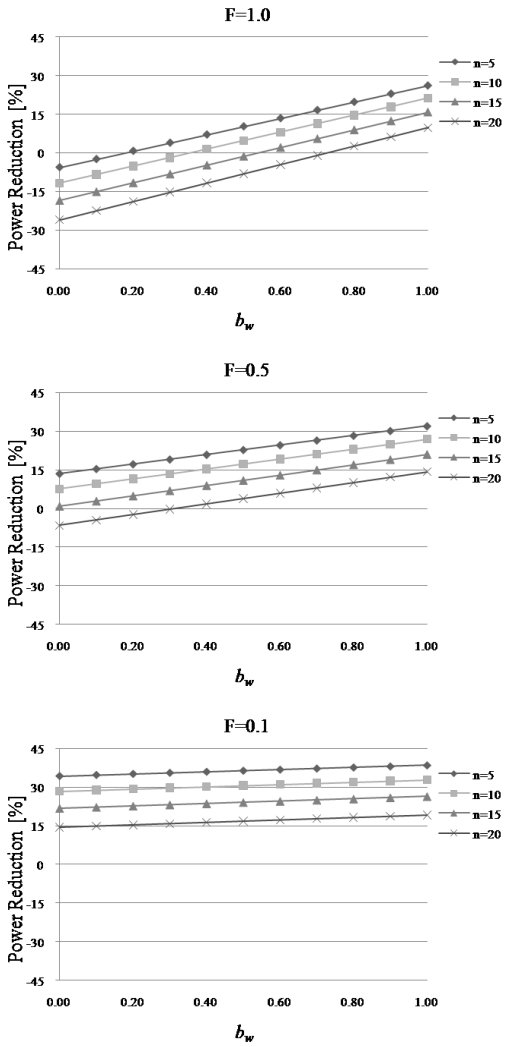

図 8　メモリバッファへの書き込み可能率 b_w を変化させた場合での消費電力削減率の概算．次のパラメータは固定値としている（h_c= 0.5, h_d= 0.5, r_d= 0.5, f_r= 0.8, f_w= 0.2）

オフとして考える必要がある．

また，図 6 においては，その他のパラメータと異なり負荷の状況より省電力効果が変わってくる．高負荷では r_d が高いほど省電力効果は高く，r_d が 5% 以下ならば省電力効果は 71.4% 以上となるが，総ディスク数に対するキャッシュディスクの割合が高い場合（10% 以上）には従来方式よりも消費電力量は多くなってしまう．中負荷から低負荷においては逆に r_d が高いほど省電力効果が低くなるが，キャッシュディスクの割合が 20% まであっても省電力効果自体は有効に働く．

各パラメータのうち，最も省電力化に影響をあたえるのはディスクの回転確率 r_d であることが分かる．ディスクドライブにおける主要な電力消費の機構はスピンドルモーターであることから，これは当然の結果といえる．その他のパラメータでは h_c と h_d は省電力に与える影響はほぼ等しい．b_w が多少劣るのは，アクセス頻度の設定によるものと考えられる．今回の概算では読み込み処理と書き込み処理の比率は 8:2 としている．

4.5　提案手法と MAID の省電力効果の比較

図 9 では，提案手法と MAID における電力消費量の削減率の見積もりとその比較をおこなった結果を示している．両手法において，n=10，m=90，f_r=0.7，f_w=0.2 を共通の固定値としており，ディスクの回転確率 r_d が 0.1，0.5，0.9 の 3 パターンに対してキャッシュディスクのヒット率 h_d を 0.7 から 0.99 まで 0.05 間隔で変化させている．なお，MAID の概算式では用いられていないキャッシュメモリのヒット率 h_c とメモリバッファの書き込み可能率 b_w は，それぞれ h_c=0.5，b_w= 0.95 として計算をおこなった．

図の結果より，r_d が同じ場合においては MAID よりも提案手法の方が電力削減率は高い値を示していることが分かる．さらに，キャッシュディスクのヒット率が高い場合においても提案手法の方が MAID よりも高い電力削減率を維持している．MAID ではキャッシュディスクのヒット率 h_d の変化によって省電力効果が大きな影響を受けているの対し，提案手法では h_d が低くなったとしてもある程度の省電力効果を得ることが出来ることがわかる．

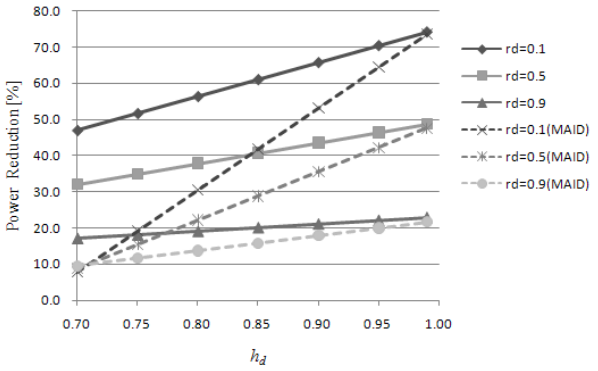

図 9　提案手法と MAID との電力量削減率の比較

5. 関 連 研 究

MAID [1] はキャッシュディスクを用いた手法を提案している．データの局所性があれば非常に良い省電力効果と良い性能維持をもたらすが，データの局所性に依存し過ぎているため局所性のない負荷によってはパフォーマンスが著しく低下してしまうという問題がある．また，MAID では省電力化とパフォーマンスの維持にのみ着目しており，信頼性については考慮されていない．

DRPM [3] はディスクドライブの回転数（RPM）を負荷に応じて動的に変更することにより，省電力化と性能の維持を実現する．しかし技術的な課題も多く，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

本研究では，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により，データ配置を有効活用して省電力化，性能の維持，および信頼性の確保を実現する手法の提案を行った．提案手法についてその省電力効果を検証するために，ストレージシステムの消費電力量を見積もる概算式を構築し，様々な状況における消費電力の見積もりを行った．見積もりの結果，アクセス時におけるディスクの回転確率が大きく省電力化には影響していることが確認出来た．

今後の課題としては，シミュレーションによるより詳細な省電力効果の検証を行い，概算式のみでは評価出来なかったパフォーマンスに関して検証を行うことがあげられる．また，メタデータの管理方法についての詳細検討や，本提案手法の信頼性についての定量的な評価を行う必要がある．

さらに，多重化構成にした場合における省電力効果とパフォーマンスへの影響および信頼性の向上についての検証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）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] E.Otto, A.Pinar, D.Rotem, and S.C.Tsao. A file allocation strategy for energy-efficient disk storage systems, 2008.

[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] 平井遥, 星野喬, 合田和夫, 喜連川優. データベースシステムにおけるプロアクティブなディスクアレイ省電力化手法に関する一考察, *DEWS*, 2008.

[6] 杉浦利之. データセンターにおける電力供給システムと省電力化, *In tutorial, SACSIS*, 2008.