複数問合せ処理を意識したディスクストレージ省電力化に関する一考察

合田　和生† 　QU　Wenyu† 　喜連川　優†

† 東京大学 生産技術研究所　〒 153-8503 東京都目黒区駒場 4-6-1
E-mail: †{kgoda,quwenyo,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　データセンタにおいて消費される電気エネルギーは急速に増大している．殊に，多数のディスクドライブが稼働するデータセンタでは，ディスクストレージは主要なエネルギー消費源であり，当該システムの省電力化は重要な研究課題である．本論文は，データベースシステムを意識したディスクストレージの省エネルギー化のアプローチを議論する．データベースシステムが有する高レベルのソウトウェア実行情報を活用する提案である，問合せ実行計画に基づくプロアクティブなディスクストレージのエネルギー制御を示すと共に，当該方式を複数の問合せが並行して処理される環境へ拡張するために，問合せ処理の遅延化可能性に基づく実行時スケジューリングを示す．シミュレーション環境における実験においては，ディスクストレージの消費エネルギーのうち，20-55%を削減する可能性が示されている．

キーワード　ディスクドライブ，省エネルギー化，グリーン，問合せ処理，ストレージ

A study on multi-query processing aware energy reduction techniques of disk storage systems

Kazuo GODA†, Wenyu QU†, and Masaru KITSUREGAWA†

† Institute of Industrial Science, The University of Tokyo　Komaba 4-6-1, Meguro-ku, Tokyo, 153-8505 Japan
E-mail: †{kgoda,quwenyo,kitsure}@tkl.iis.u-tokyo.ac.jp

Abstract　Electric energy consumed in data centers is rapidly growing. Disk storage is a non-negligible energy consumer. Rather, in light of recent data-intensive systems where a number of disk drives are incorporated, the disk storage may be what we must consider primarily. Energy saving of such disk storage is a grand challenge for IT research and development. The paper discuss our new database server assisted approach of saving disk energy consumption. First, we study a proactive energy control technique based on query execution plans, which should be seen as an approach of exploiting high-level software behavioral information that the database server holds. Second, we propose a query processing scheduling technique based on deferability of the queries so that the proactive energy control could be extended to multiple query processing environments. The simulation experiments shows that 20-55% energy of the disk storage can potentially saved.

Key words　Disk Drive, Energy Saving, Green, Query Processing, Storage

1. はじめに

データセンタにおいて消費される電気エネルギーは年率25%で急速に増大しており [9]，2009 年には電力コストはサーバの調達コストの 2 倍に達すると予測されている [3]．より多くの冷却システムと給電装置がデータセンタには備えられるようになっており，典型的なデータセンタでは TCO の 44%が電気エネルギーと関連装置によって消費されるに至っている [2]．コスト上の問題に加えて，エネルギーと熱の管理はデータセンタの設計と運用の鍵となっている．増大するエネルギー消費ま，今後の IT システムの設計空間を著しく制限する可能性があり [34]，省エネルギー化は重要な研究開発課題である．

IT システムの中で，ストレージシステムの消費エネルギーは無視できない [31]．現時点で，データセンタでは，ストレージシステムによって約 27%のエネルギーが消費されているとされ [23]，デジタル情報が爆発的に増大し，膨大な記憶管理資源がストレージシステムには組み込まれるようになっており，特にデータインテンシブな IT システムにおいては，その消費エネルギーはますます増加する可能性がある．Q. Zhu らの報告では，大規模なオンライントランザクション処理システムにお

いては，約 71%の消費エネルギーがディスクドライブに起因しているとされる [38]．ディスクストレージの消費エネルギーの削減は，サーバのプロセッサやネットワーク装置と並んで，不可欠のものと言えよう．

ディスクドライブの主様なエネルギー消費源はスピンドルモータである(注1)．消費エネルギーを削減するためには，適時にスピンドルモータを停止し，また，駆動させるのが自然なアプローチである．しかし，ディスクドライブは機械的制御がなされるものであり，スピンドルモータの停止と駆動による制御損失は無視できない．スピンドルモータを停止させた直後に，アクセス要求が到来した場合，制御損失によって逆に消費エネルギーを増加させることとなる．スピンドルモータの停止と駆動を行う時間を決定することが，ディスクストレージの省エネルギー化の鍵である．

本論文は，データベースシステムを意識したディスクストレージの省エネルギー化のアプローチを議論する．即ち，データベースシステムが有する高レベルのソウトウェア実行情報を活用することにより，ディスクストレージの消費エネルギーを削減することを提案する．著者らは論文 [41] において，単一問合せ処理環境における基礎的な実験結果を示したが，本論文では複数問合せ処理環境ににこれを拡張する．また，問合せ処理を積極的に遅延させる発見的なスケジューリング手法を提案し，シミュレーション環境における実験結果によって有効性を検証する．

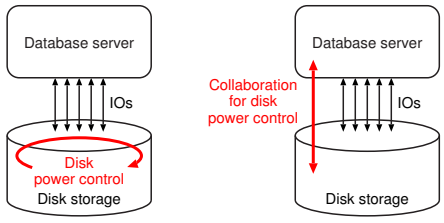
本論文は以下の通り構成される．2 章では，商用ディスクドライブのエネルギー消費を簡潔に示し，著者らの動機を述べる．3 章では，ディスクストレージのエネルギー消費を意識した問合せ処理手法の基本的な発想を示し，4 章では，ディスクドライブシミュレータを用いた実験による提案手法の評価を示す．5 章で関連研究をまとめるとともに，6 章で論文を結ぶ．

2. ディスクドライブの電気的エネルギー消費

多くの商用ディスクドライブは**スタンバイ**状態へ移行する機能を有している(注2)．一般に，スタンバイ状態のディスクドライブはヘッドがプラッタからアンロードされており，スピンドルモータの回転は停止している．このため，入出力アクセスが存在する**アクティブ**状態や即座にアクセスが可能な**アイドル**状態と比較して，消費するエネルギーは格段に小さい．しかし，スタンバイ状態への移行とそれからの復帰には時間損とエネルギー損を伴う．特に，スピンアップ (スタンバイ状態からアイドル状態への復帰) には，数秒から数十秒の時間と，数百ジュールのエネルギーが掛かる．これらの値は，プロセッサ，メモリ及びネットワーク装置など他のコンピュータ部品には見られない顕著なものである．

(注1)：ディスクドライブのスピンドルモータは空気抵抗に逆らって高速でプラッタを回転させる．モータの消費電力は，回転速度の 3 乗から 5 乗に比例するとされる [1]．

(注2)：本論文は，典型的なスタンバイモード，即ち，ヘッドがランプにアンロードされ，スピンドルモータが停止した状態を議論する．最近の商用ディスクストレージの中には，ヘッドアンロードモードや低回転速モードなどの低エネルギー消費モードを有するものがある [16]．著者らのアプローチは，容易にこれらの低エネルギー消費モードに応用することが可能である．



(a) Conventional strategy　(b) Proactive strategy

図 1　ディスクストレージのエネルギー制御の比較．
Fig. 1　Comparison of disk storage energy controls.

スタンバイ状態へ移行することによる削減エネルギーと，スピンダウン (アイドル状態からスタンバイ状態への移行) とスピンアップに伴うエネルギー損が拮抗するアイドル時間を**ブレークイーブン**時間と称する．仮に将来にディスクドライブに発行される全ての入出力を正確に予測することが可能であれば，ブレークイーブン時間よりアイドル時間においてのみディスクドライブをスタンバイ状態に移行し，新たな入出力要求が到来する事前にアイドル状態へ復帰することが可能である．このような Oracle Power Management (OPM) [22] は，例えば似通ったパターンのディスクアクセスが繰り返し行われる科学技術計算などの限定されたアプリケーションでは実現可能である可能性があるもの，一般には難しく，特に非決定的な処理を特徴とするデータベースシステムへの応用は不可能である．

ディスクストレージの省エネルギー化については，別のアプローチが取られて来た．多くの従来研究では，ディスクドライブのエネルギー状態の制御とサーバ上のソフトウェアコードの振舞は直接的には連携せず，むしろ，ストレージシステム内部でアクセス統計やパターン学習に基づいて行われて来た．図 1(a) に図示するこのようなアプローチは，実装がインターオペーラビリティの確保が容易である利点を有するが，著しい省エネルギー化が期待できるとは言いがたい．

著者らは，これらの従来研究とは異なる学術的研究として，図 1(b) に示すような高レベルのデータベース処理情報をディスクストレージのエネルギー管理に活用することを提案している．上位層の有するソフトウェア実行情報をストレージシステムが獲得することにより，将来の入出力を直接予測することが可能となり，OPM に近い省エネルギー効果を得ることが可能となる．著者らは論文 [41] において，問合せ実行計画をディスクストレージのエネルギー管理に活用することを議論した．問合せ処理の事前に生成される問合せ実行計画は，スピンドルモータの停止と駆動を判断するための大きなヒントとなり得る．このようなプロアクティブな手法により，より高い省エネルギー化効果が期待される．

3. データベースの問合せ処理とディスクストレージのエネルギー制御の連携

本論文では，アドホック問合せを想定して議論する．アドホック問合せは大規模情報の解析などに用いられ，一般にその実行時間は長いが，その間，全てのディスクドライブは必ずし

```
ProcessQuery(Query q)
{
  QueryPlan p = GenerateQueryPlan(p);
  Step s = p.firststep();
  do {
    Time t = PredictStepTime(s);
    Disks di = RetrieveIdleDisks(s);
    /*
     * Spin down idle disks if the current step is
     * longer than break-even time.
     */
    if(t > di.BreakEvenTime)
      SpinDown(di);
    /*
     * Set timer to spin up disks that are active
     * at the next step in advance.
     */
    if(s.nextstep()){
      Disks da = RetrieveActiveDisks(s.nextstep());
      if(t > da.SpinUpTime)
        SetTimer(t - da.SpinUpTime, SpinUp, da);
    }
    /*
     * Execute the current step.
     */
    ExecuteStep(s);
  } while(s = s.nextstep());
}
```

図 2　プロアクティブなエネルギー制御の疑似コード.
Fig. 2　A pseudocode of proactive energy control.

もアクティブであるわけでなく，一部のディスクは時にアイドルである場合がある．著者らはこの特性に着目し，当該アイドル時間を問合せ実行計画から予測し，プロアクティブにディスクドライブをスタンバイ状態へ移行させることを提案する．さらに，当該技法を複数の問合せが並行して処理される環境へ拡張し，消費エネルギーを意識した実行時スケジューリングを議論する．また，発見的な技法として，遅延化問合せ処理を示す．

3.1　問合せ実行計画に基づくプロアクティブなディスクストレージのエネルギー制御

データベースサーバに問合せが到着すると，問合せは先ず問合せ実行計画へ変換され，当該実行計画に基づき問合せが処理される．実行計画は，通常，1 つ以上のステップから構成され，各ステップ毎に関係表と索引へのアクセスが決定されている．このような問合せ実行計画を解析することにより，各ディスクがいつアクセスされ，いつアイドルとなるのかを容易に予測することが可能となる．

例えば，データベースが 2 つの関係表 R と S をそれぞれボリューム V1 と V2 に有しているとする．R と S を結合する問合せが与えられ，ハッシュ結合に基づく問合せ実行計画が生成されたとする．即ち，問合せは 2 つのステップで行われ，まず，関係表 R を走査しハッシュ表を主記憶上に生成し，その後，関係表 S を走査し当該関係表を検索する．第 1 のステップでは関係表 R のみが，第 2 のステップでは関係表 S のみがアクセスされるため，第 1 のステップの実行中にボリューム V1 をスタンバイ状態へと移行し，第 2 のステップの実行中にボリューム V2 を同様にすることは，省エネルギー化に有益なアプローチである．

問合せ処理の既存のソフトウェアコードに若干の改変を加えることにより，このような問合せ実行計画に基づくプロアクティブな技法を実現することが可能となる．図 2 に疑似コードを示す．問合せ処理機構は，問合せの各ステップの冒頭において，当該ステップがブレークイーブン時間より長いと予測される場合に，当該ステップ中にアクセスされないディスクドライブにスピンダウンコマンドを送付する．また，各ステップの終盤においては，次のステップでアクセスされるであろうディスクドライブに事前にスピンアップコマンドを発行する．このようなプロアクティブ戦略の利点を以下にまとめる．

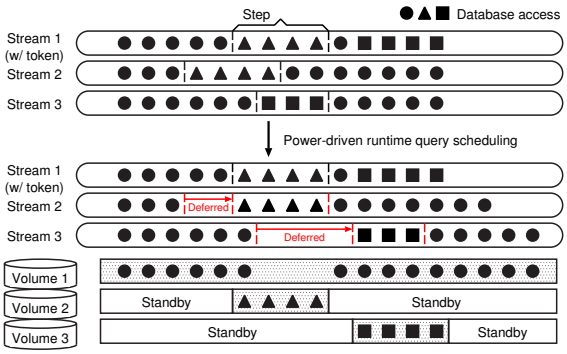


図 3　遅延化問合せ処理.
Fig. 3　Deferred query processing.

- ステップが始まると，当該ステップ中にアクセスされないディスクドライブは即座にスタンバイ状態へと移行するため，ステップ実行中のアイドルなディスクドライブの消費電力を最小化することが可能となる．
- ブレークイーブン時間に基づいてディスクドライブのスピンダウンが決定されるため，エネルギー制御によって大きな制御損を被ることを避けることができる．

通常，スピンドルモータの制御には数秒から数十秒を要するため，上記のプロアクティブな戦略は，主にアドホック問合せなどの数分から時に数時間を要する処理に有効である．

3.2　問合せ処理の遅延化可能性に基づく実行時スケジューリング

本論文では，このようなプロアクティブなディスクストレージのエネルギー制御を，複数の問合せが並行して処理される環境に拡張する．この際，データベースサーバは，複数の問合せを調停することにより消費エネルギーを調整する必要があり，その有効な手法として，遅延化問合せ処理を提案する．

図 3 に遅延化問合せ処理による省エネルギー指向の実行時スケジューリングを示す．複数の問合せ系列がデータベースサーバに与えられているとする．この際，ディスクドライブをスピンアップする権限である**スピンアップトークン**を導入する．当該トークンを有する系列のみがディスクドライブを必要に応じてスピンアップすることが可能であり，トークンを有さない系列が問合せを処理するためにスピンダウンされたディスクドライブをアクセスする際には，当該ドライブが他の系列によってスピンアップされるまで待つ必要がある．付与されるトークンの総数を小さく設定することにより，複数の問合せ系列をディスクドライブのエネルギー状態に基づき調停することが可能となる．なお，問合せ系列間の公平性のために，系列は一定時間トークンを保有した場合には当該トークンを他の系列に譲り渡すものとする．

表 1　ディスクドライブのシミュレーションパラメータ.
Table 1　Disk drive simulation parameters.

Model	IBM Ultrastar 36Z15
Capcity	18.4 GB
Rotational speed	15000 rpm
Avg. seek time	3.4 ms
Transfer rate	55.0 MB/s
Active power	39.0 W
Idle power	22.3 W
Stand-by power	4.15 W
Spin-down time	15.0 s
Spin-down energy	62.25 J
Spin-up time	26.0 s
Spin-up energy	904.8 J
Idleness threshold	60.0 s

4.　シミュレーション実験による有効性の検証

著者らは，前節で議論したプロアクティブなエネルギー制御技法ならびに遅延化問合せ処理を，トレース駆動のシミュレーション環境と TPC-H ベンチマークを用いて評価した．本節では当該実験による有効性の検証を示す.

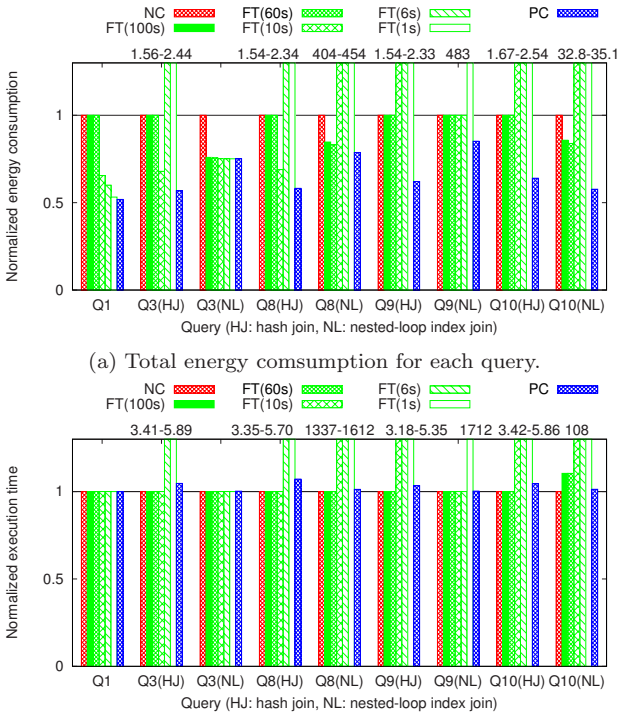
4.1　シミュレーション実験環境

評価実験では著者らの開発したディスクストレージシミュレータを実験基盤として用いた．DiskSim [4] ならびにその高度化版 [26,37] と同様に，当該シミュレータは，入出力に基づきイベント駆動のシミュレーションを行い，実行時間や消費エネルギーなどの結果を返す．当該シミュレータ上で，さらに，著者らは提案手法であるプロアクティブなエネルギー制御技法ならびに遅延化問合せ処理を模擬できるように改造を行った．即ち，シミュレータに入力する入出力トレースに，問合せと問合せ実行計画の情報を付与することにより，入出力に焦点を絞った問合せ処理の模擬を行った.

表 1 に実験で用いたディスクドライブのモデルパラメータを示す．当該モデルパラメータは，必ずしも最新の製品ではない IBM Ultrastar 36Z15 に基づいているものの，多くの研究 [5,19,28–30,36,38–40] で用いられている利点がある.

2 つの Xeon プロセッサと 2GB の主記憶を有する Linux サーバ上で，HiRDB [20] を用いて TPC-H ベンチマークの問合せを実行した．この際，問合せ実行計画を記録するとともに，カーネルレベルの入出力トレーサを用いて各ステップ毎の入出力動作 (読込みならびに書込みの要求と完了) を記録した.

シミュレーション実験においては，データベースは 3 つのボリューム V1，V2 及び V3 から構成され，各ボリュームはパリティを用いないで複数のディスクドライブにストライプされるものとした．大きい関係表である LINEITEM と ORDER はそれぞれ V1 と V2 に格納され，それ以外の関係表は V3 に格納されるものとした．このようなデータベースの編成は，TPC によって公開されている結果情報においても，しばしば見られるものである.



(a) Total energy comsumption for each query.

(b) Execution time of each query.

図 4　各種問合せ処理方式におけるエネルギー削減と時間損.
Fig. 4　Enegy saving and time penalty for different types of query processing.

4.2　単一問合せ処理環境

データベースサーバにおいて単一のアドホック問合せが処理される環境下で，プロアクティブなディスクストレージのエネルギー制御の有効性を評価した.

スケールファクタ 2 のデータセットを用い，シミュレーション環境において各問合せの処理に掛かる実行時間と消費エネルギーを測定した．Q1 は単純に関係表を走査する問合せであり，Q3，Q8，Q9 および Q10 は複数の結合を行うより複雑な問合せである．評価の多様性のため，ハッシュ結合と索引結合の 2 つの結合方式のそれぞれを用いた場合を評価した．この際，**NC**，**FC** 及び **PC** なる 3 つのエネルギー制御方式を比較した．**NC** はディスクストレージのエネルギー制御を行わない場合である．**FT** はアイドル時間の閾値に基づくエネルギー制御を意味し，閾値は 1 秒から 100 秒まで変化させて計測した．最後に，**PC** が提案手法であるプロアクティブなエネルギー制御を意味する.

図 4 に計測結果を示す．なお，ここで消費エネルギーとは，問合せ処理中の平均的な電力消費量ではなく，問合せ実行中に掛かる総消費エネルギー量であることに注意されたい．実行時間ならびに消費エネルギーは，**NC** を基準として正規化されている．まず最初に，従来方式である **FT** について検証したい．省エネルギー効果を高めるためには，より小さい閾値を設定することが一見望ましい．Q1 のような問合せに関しては有効であり，**PC** と同等の省エネルギー効果を得ている．しかし，その他の問合せにおいては，著しい実行時間の長期化と消費エネルギーの増大を生じる場合がある．例えば，Q8(NL) において

は，7 つの関係表が索引を用いて結合され，3 つのボリュームに渡って入出力アクセスが発行される．LINEITEM を有する V1 と ORDERS を有する V2 は頻繁にアクセスされるものの，V3 のデータはデータベースキャッシュに乗っている場合が多く，アクセスされる頻度は比較的低い．小さい閾値設定の場合，V3 は稀にスピンダウンされることがあるが，その後に入出力が到来し，再度スピンアップされる．スピンアップには数十秒を要するため，V3 がスピンアップしている間，以降の問合せ処理はブロックされる．すると，今度は逆に V1 と V2 に入出力が到着しないことになり，当該ボリュームがスピンダウンされる場合があり，大きな制御損に至った．一方，一方，プロアクティブなエネルギー制御である **PC** は，全ての問合せに対して最大の省エネルギー効果を得ており，概ね 15-50%の省エネルギー化を達成している一方，観測された性能損は小さい．

4.3 複数問合せ処理環境

データベースサーバにおいて複数のアドホック問合せが処理される環境下で，プロアクティブなディスクストレージのエネルギー制御と問合せ処理を遅延させる実行時スケジューリングの有効性を評価した．

同様なシミュレーション環境において，1 つ以上の問合せ系列が並行して処理される場合について，実行時間と消費エネルギーを計測した．この際，スケールファクタ 2 と 20 のデータセットをそれぞれ用意した．各問合せ系列では Q1 から Q10 のアドホック問合せがランダムに要求されるものとし，並行処理される系列の数を 1 から 50 まで変化させ，全問合せ処理にディスクストレージが必要とするエネルギーを計測した．この際，**NC**，**FT**，**PC** 及び **PC++**の 4 つのケースを比較した．++は遅延化問合せ処理の有効化を意味する．なお，**FT** における閾値は 60 秒とし，**PC** におけるトークンは 1 個に制限した．

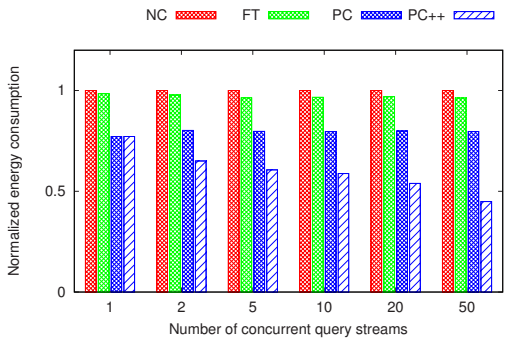
図 5 に結果を示す．**FT** による省エネルギー化は高々5-10%程度であるのに対し，**PC** は 20%程度の省エネルギー化を達成しており，プロアクティブなエネルギー制御の有用性は明らかである．更に **PC++**では利得を拡大し，40-55%の省エネルギー化を達成しており，他のいずれの手法にも優る結果を得た．プロアクティブなエネルギー制御と遅延化問合せ処理を併用することによる，ディスクストレージの著しい省エネルギー化の可能性が示されたと言えよう．

5. 関 連 研 究

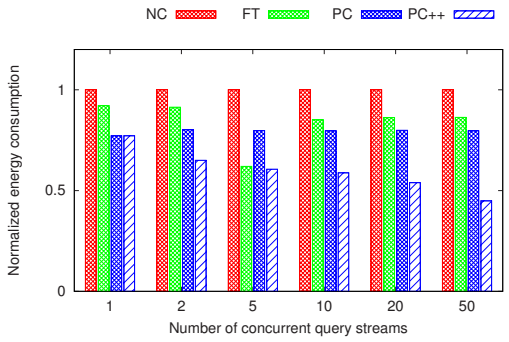
これまでにも，幾つかのディスクストレージの省エネルギー化を目指すが発表されてきた．本論文ではこれらを 6 つのアプローチに分類に分類して，簡潔に示す．

a） アイドル時間の閾値

閾値に基づきスピンダウンを行うアプローチは活発に研究されている．最も簡易な手法は，一定のアイドル時間の閾値に基づき，ディスクドライブを省エネルギーモードへ遷移させるものである．即ち，最後のディスクアクセスから与えられた閾値時間が経過すると，ディスクドライブをスピンダウンするものである．このような単純な技法は多くの商用ディスクドライブに採用されている．また，より高度な手法として，閾値を適応



(a) Scale factor: 2.0



(b) Scale factor: 20.0

図 5 複数のアドホック問合せ処理におけるエネルギー削減．
Fig. 5 Energy saving on multiple ad-hoc query execution.

的に制御する技法も提案されている [8,11,15,17]．このような閾値に基づく技法は，主に対話的なアプリケーションが実行され，ユーザがディスクのエネルギー制御によって生じる時間損をある程度許容することができるバッテリー駆動のモバイルコンピューティングやラップトップコンピュータ環境においては有効に機能するものである．しかし，データセンタで頻繁に見られるようなデータインテンシブなシステムにこれらを適用することは難しいと思われる．

b） アクセスの局所性

2 つめのアプローチは，ディスクストレージへのアクセスの局所性を活用しようとするものである．Massive Array of Idle Disks (MAID) [6,7] なるストレージシステムは，キャッシュボリュームとキャッシュボリュームから構成される．頻繁にアクセスがなされるブロックをキャッシュボリュームに積極的に複製することにより，MAID ではメインボリュームに長いアイドル時間を生成し，これによってメインボリュームをスピンダウンする機会を得ようとする．MAID のアプローチは，アーカイブ用途のストレージシステムとして商用化されるに至っている [24]．また，Popular Data Concentration (PDC) [5,27] は，異なる種類のディスクドライブを間でブロックを移送するという異なる発想である．PDC では，頻繁にアクセスされるブロックを高速なディスクドライブに配置し，そうでないブロックを低速なブロックに配置することにより，システム全体での省エネルギー化を達成しようとする．アクセスの局所性を活用するこれらの技法は，主にファイルサーバ等において有効に機能するものである．

c） 多段速のディスクドライブ

回転速度の変更が可能な多段速のディスクドライブについては，盛んに研究が行われている．Dynamic RPM (DRPM) [12,13] は，ディスクドライブの回転速度を動的に制御する発想である．提案されたアルゴリズムでは，直近に処理された入出力の応答時間を記録し，消費エネルギーと性能をバランスさせるような回転スピードを決定する．また，Hibernator [38] は PDC と類似の提案であるが，多段速のディスクドライブに着目したものである．即ち，Hibernator は，ストレージシステムを回転速度に応じた複数のティアに分割し，ティア間でブロックを動的に移送するとともに，各ティアにおける回転速度をアクセス頻度に基づき定期的に変更する．これらの多段速のディスクドライブを用いる試みは一見有力なアプローチであるが，このような多段速のディスクドライブは試作機が報告されているに過ぎず [25,35]，なおも現時点において製品として出荷されてはいない．

d） 記憶空間の冗長性

殆んどのディスクストレージは RAID，即ち可用性を高めるために記憶空間に冗長性を持たせる機能を有している．D. Li らは，従来型の RAID 編成の上で省エネルギー化を行う Energy Efficient RAID (EERAID) と名付けた入出力スケジューリングとキャッシュ管理の技法を提案している [18]．その基本的な発想は，RAID 制御器がパリティやミラー情報を更新するために発行する入出力を集約するとともに，スピンダウン中のディスクドライブに本来格納されているブロックのエビクションをなるべく避けるものである．EERAID は，元々，多段速のディスクドライブを念頭に提案されていたが，当該手法は通常のディスクドライブ向けに拡張されている [19]．一方，RIMAC [36] は類似のキャッシュ管理を提案しているが，2 レベル構成のキャッシュとディスクドライブの連携によって，入出力要求を変換することに焦点を絞っている．記憶空間の冗長性を活用するこれらのアプローチについては，E. Pinheiro らによって解析的なモデルが導入され，検証が行われている [28]．

一方，Power-Aware RAID (PARAID) [32] は異なる従来型の RAID-5 編成を変更するものである．非対称のパリティ配置を行うことにより，アクティブ状態のディスクドライブの数を動的に変更することが可能となる．このようなギア変換によってデータセンタで頻繁に見られるような急峻な負荷変動へ適応することを特徴としている．

e） キャッシュ管理

ストレージシステムにはより広大なキャッシュメモリ空間が導入されており，これらのキャッシュメモリを省エネルギー化に活用することは自然なアプローチであろう．A. Papathanasiou らは，強いバースト性と長いアイドル時間と有するディスクアクセスパターンを生成するための先読みとキャッシュ管理の技法を提案している [26]．また，Q. Zhu らは，Online Power-aware Greedy algorithm (OPG) と称する省エネルギー型のキャッシュ交換アルゴリズムを提案するとともに，様々なキャッシュ管理手法の評価を報告している [39,40]．

f） アプリケーションとの連携

アプリケーションとの連携によってディスクストレージの消費エネルギーを削減する試みが報告されている．協調的入出力 Cooperative IO (Coop-IO) [33] は省エネルギー化のための入出力システムコールの提案である．これにより，ユーザは各入出力に対し遅延可能性 (deferability) と中断可能性 (abortability) を指定することができる．協調的な入出力，即ち，遅延可能かもしくは中断可能な入出力は，積極的に遅延もしくは中断され，これにより入出力系列を集約し，より長時間のアイドルを生成する．一方，Y. Lu も同時期に同様の入出力要求スケジューリングを提案しているが，入出力を遅延させるのではなく，より先行して処理する点に特徴を有する [21]．

一方，T. Heath らは，入出力要求を集約するために，ソースコードレベルで入出力命令を調整するコンパイラに基づくアプリケーション変換を提案している [14]．S. Son らは，科学技術計算アプリケーションに焦点を絞った同様のソフトウェア指向の技法を示している [29,30]．Son らのコンパイラはソースコードを解析してループ計算を再構築するとともに，省エネルギー化のためのファイルレイアウトを調整する．C. Gniady らは，Program-Counter Access Predictor (PCAP) と称する技法を提案している [10]．即ち，実行中のアプリケーションのプログラムカウンタを観察することにより，入出力アクセスパターンを学習し，これに基づき適応的にディスクのエネルギーモードを制御する．

g） 提案手法との比較

ディスクストレージの省エネルギー化に関する著者らのアプローチは，データベースシステムとの強い連携によるところに特徴を有する．上記で最後のアプローチに示した関連研究と一見似通っているかもしれないが，以下のような相異点がある．

PCAP は大変興味深い発想であるが，ディスクドライブのエネルギー制御の基準とする情報が著しく限定されている．C. Gniady らはむしろ省エネルギー化を達成するために必要とする情報を最小化することに注力しており，将来のアクセスパターンを直接示す情報は利用されていない．これに対して，著者らの提案するプロアクティブなアプローチでは，問合せ実行計画なる上位層の豊富な情報を活用し，学習過程を経ず，直接的にエネルギーモードを制御する．

著者は，cooperative IO が究極の発想であると考えている．仮に全てのアプリケーションのソフトウェアコードを書き換えることが可能であれば，当該機構は著しい影響力を有するであろう．しかし，現時点において省エネルギー化のためにソースコードを敢えて改変することは限られたアプリケーションでのみ受け入れられるものである．特に業務アプリケーションの全てのソースコードに当該機構を組み込むことは現実的ではない．一方，Heath と Son の華麗な発想は，類似のデータアクセスがループ内で繰り返される科学技術計算アプリケーションにおいては有力なアプローチであろう．しかし，非決定的な振舞を特徴とするデータベースシステムではに当該アプローチを適用すること困難である．本論文では，業務用途で広く利用されているデータベースシステムに焦点を絞った提案を行っている．若

干の改変がデータベース管理システムには必要となるが，個々のアプリケーションのソフトウェアコードはそのまま利用が可能である．即ち，データベースシステムにおける新たな省エネルギー化手法を提案するものである．

6. おわりに

本論文は，データベースシステムを意識したディスクストレージの省エネルギー化のアプローチを議論した．即ち，データベースシステムが有する高レベルのソウトウェア実行情報を活用することにより，ディスクストレージの消費エネルギーを削減することを提案した．論文 [41] において示した方式を複数問合せ処理環境に拡張し，問合せ処理を積極的に遅延させる発見的なスケジューリング手法を提案した．シミュレーション環境における実験においては，ディスクストレージの消費エネルギーについて，単一のアドホック問合せ環境においてプロアクティブなエネルギー制御を用いることにより 20-50%を削減する結果を得た．また，複数問合せ環境においてはプロアクティブなエネルギー制御と遅延化問合せ処理を併用することにより，40-55%を削減する可能性が示された．

本論文では，ディスク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エネルギーに焦点を絞って議論した．今後，交直変換器や RAID 制御器，キャッシュメモリなどの他の構成部品に関しても考慮した総合的な解析により，アプローチの有用性を検証したい．

謝　辞

本研究の一部は，文部科学省リーディングプロジェクト e-Society 基盤ソフトウェアの総合開発「先進的なストレージ技術」の助成により行われた．協力企業である株式会社日立製作所より多くの有益なコメントを頂戴した．感謝する次第である．

文　献

[1] D. Anderson and W. Whittington. Hard Drives: Today & Tomorrow. In *Tutorial, USENIX Conf. on File and Storage Tech.*, 2007.

[2] APC. Determining Total Cost of Ownership for Data Center and Network Room Infrastructure. White paper, 2002.

[3] B. Rudolph. Storage In an Age of Invonvenient Truths. SNW2007Spring, 2007.

[4] J. S. Bucy and G. R. Ganger. The disksim simulation environment: Version 3.0 reference manual. Online manual available at `http://www.pdl.cmu.edu/DiskSim/`, 2003.

[5] E. V. Carrera, E. Pinheiro, and R. Bianchini. Conserving Disk Energy in Network Servers. In *Proc. Int'l Conf. on Supercomputing*, pages 86–97, 2003.

[6] D. Colarelli and D. Grunwald. Massive Arrays of Idle Disks for Storage Archive. In *Proc. ACM/IEEE Conf. on Supercomputing*, pages 1–11, 2002.

[7] D. Colarelli, D. Grunwald, and M. Neufeld. The Case for Massic Arrays of Idle Disks (MAID). In *Proc. USENIX Conf. on File and Storage Tech.*, 2002.

[8] F. Douglis, P. Krishnan, and B. Bershad. Adaptive disk spin-down policies for mobile computers. In *Proc. USENIX Symp. on Mobile and Location-Independent Computing*, pages 121–137, 1995.

[9] F. Moore. More power needed. Energy User News, 2002.

[10] C. Gniady, Y. C. Hu, and Y-H. Lu. Program Counter-Based Prediction Techniques for Dynamic Power Management. *IEEE Trans. Comput.*, 55(6):641–658, 2006.

[11] R. A. Golding, P. Bosch, C. Staelin, T. Sullivan, and J.Wilkes. Idleness is not sloth. In *Proc. USENIX Tech. Conf.*, pages 201–212, 1995.

[12] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke. DRPM: Dynamic Speed Control for Power Management in Server Class Disks. In *Proc. Int'l Symp. on Comput. Arch.*, 2003.

[13] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke. Reducing Disk Power Consumption in Servers with DRPM. *IEEE Computer*, 36(12):59–66, 2003.

[14] T. Heath, E. Pinheiro, J. Hom, U. Kremer, and R. Bianchini. Application Transformations for Energy and Power-Aware Device Management. In *Proc. Int'l Conf. on Parallel Arch. and Compilation Tech.*, pages 121–130, 2002.

[15] D. P. Helmbold, D. D. E. Long, T. L. Sconyers, and B. Sherrod. Adaptive disk spin-down for mobile computers. *Mobile Networks and Applications*, 5(4):285–297, 2000.

[16] HGST Inc. Quietly cool. White Paper, HGST, 2004.

[17] P. Krishnan, P. M. Long, and J. S. Vitter. Adaptive disk spindown via optimal rent-to-buy in probabilistic environments. In *Int'l Conf. on Machine Learning*, pages 233–330, 1995.

[18] D. Li and J. Wang. EERAID: Energy Efficient Redundant and Inexpensive Disk Array. In *Proc. ACM SIGOPS Euro. Workshop*, 2002.

[19] D. Li, J. Wang, and P. Varman. Conserving Energy in Conventional Disk based RAID Systems. In *Proc. Int'l Workshop on Storage Network Arch. and Parallel I/Os*, pages 65–72, 2005.

[20] Hitachi Ltd. Hitachi Relational Database Management System Solutions for Disaster Recovery to Support Business Continuity. Review Special Issue, Hitachi Technology, 2004.

[21] Y. Lu, L. Benini, and G. Micheli. Power-aware operating systems for interactive systems. *IEEE Trans. Very Large Scale Integration Syst.*, 10(2):119–134, 2002.

[22] Y. Lu and G. Micheli. Comparing system-level power management policies. *IEEE Design and Test of Comput.*, 18(2):10–19, 2001.

[23] Maximum Throughput Inc. Power, heat, and sledgehammer. White paper, 2002.

[24] F. Moore and A. Guha. Introducing COPAN Systems MAID Architecture (Massive Array of Idle Disks). White Paper, Copan Systems, 2004.

[25] K. Okada, N. Kojima, and K. Yamashita. A Novel Drive Architecture of HDD: Multimode Hard Disc Drive. In *Proc. Int'l Conf. on Consumer Electronics*, pages 2213–2215, 2000.

[26] A. E. Papathanasiou and M. L. Scott. Energy Efficient Prefetching and Caching. In *Proc. USENIX Tech. Conf.*, 2004.

[27] E. Pinheiro and R. Bianchini. Energy Conservation Techniques for Disk Array-Based Servers. In *Proc. Int'l Conf. on Supercomputing*, pages 68–78, 2004.

[28] E. Pinheiro, R. Bianchini, and C. Dubnicki. Exploiting Redundancy to Conserve Energy in Storage Systems. In *Proc. ACM SIGMETRICS Conf.*, pages 15–26, 2006.

[29] S. W. Son, G. Chen, and M. Mandemir. Disk Layout Optimization for Reducing Energy Consumption. In *Proc. Int'l Conf. on Supercomputing*, pages 274–283, 2005.

[30] S. W. Son, M. Mandemir, and A. Choudhary. Software-Directed Disk Power Management for Scientific Applications. In *Proc. IEEE Parallel and Distributed Processing Symp.*, page 4b, 2005.

[31] T.C. 9.9. *Datacom Equipment Power Trends and Cooling Applications*. ASHRAE, 2005.

[32] C. Weddle, M. Oldham, J. Qian, A. A. Wang, P. Reiher,

and G. Kuenning. PARAID: A Gear-Shifting Power-Aware RAID. In *Proc. USENIX Conf. on File and Storage Tech.*, pages 245–260, 2007.

[33] A. Weissel, B. Beutel, and F. Bellosa. Cooperative I/O – A Novel I/O Semantics for Energy-Aware Applications. In *Proc. USENIX Symp. on Operating Syst. Design and Imple.*, pages 117–130, 2002.

[34] S. Worth. Green Storage. SNIA Education, 2006.

[35] H. Yada, H. Ishioka, T. Yamakoshi, Y. Onuki, Y. Shimano, M. Uchida, H. Kanno, and N. Hayashi. Head positioning servo and data channel for HDDs with multiple spindle speeds. *IEEE Trans. Magnetics*, 36(5):2213–2215, 2000.

[36] X. Yao and J. Wang. RIMAC: A Novel Redundancy-based Hierarchical Cache Architecture for Energy Efficient, High Performance Storage System. In *Proc. EuroSys*, pages 249–262, 2006.

[37] J. Zedlewski, S. Sobti, N. Garg, and F. Zheng. Modeling hard-Disk Power Consumption. In *Proc. USENIX Conf. on File and Storage Tech.*, pages 217–230, 2003.

[38] Q. Zhu, Z. Chen, L. Tan, Y. Zhou, K. Keeton, and J. Wikes. Hibernator: Helping Disk Arrays Sleep through the Winter. In *Proc. ACM Symp. on Operating Syst. Principles*, pages 177–190, 2004.

[39] Q. Zhu, F. M. David, C. F. Devaraj, Z. Li, Y. Zhou, and P. Cao. Reducing Energy Consumption of Disk Storage Using Power-Aware Cache Management. In *Proc. Int'l Symp. on High Performance Comput. Arch.*, pages 118–129, 2004.

[40] Q. Zhu and Y. Zhou. Power Aware Storage Cache Management. *IEEE Trans. Comput.*, 54(5):587–602, 2005.

[41] 上野 裕也, 合田 和生, and 喜連川 優. データベースシステムの問い合わせ実行計画を利用したディスクアレイ省電力化に関する一考察. **日本データベース学会** *Letters*, 6(1):85–88, 2007.