データベースシステムの問合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

A Study on Disk Array Power Reduction using Query Plan of Database Systems

上野 裕也♥ 合田 和生♦ 喜連川 優♦

Yuya UENO
Kazuo GODA Masaru KITSUREGAWA

従来，コンピュータシステムにおける消費電力に関しては，主にモバイルコンピュータや組込みコンピュータなどの資源が制限された環境におけるプロセッサの省電力化を中心に検討がなされてきた．近年では，データセンタなどの大規模システムにおいてもその消費電力が問題となりつつあり，また，プロセッサだけでなく周辺の入出力機器を含めたサブシステムの省電力化が求められるようになっている．特に，システムの管理するデータ量が急激に増大している中，ディスクドライブの省電力化は極めて重要な課題である．本論文では，多数のディスクドライブから構成されるディスクアレイの省電力化を目指し，データベースシステムの有する問合せ実行計画を利用した新しいディスクドライブの制御方式を提案する．独自の方式により構築したディスクドライブに関する消費電力モデルを示すとともに，当該モデルに基づく解析的検討により，従来方式と比較して大きな効果が得られることを明らかにする．

Power consumption of processors has been studied so far mainly in resource-limited computing environments such as mobile computers and embedded computers. Nowadays, the power consumption of large systems (e.g. data centers) is beginning to be a new problem, and other subsystems including input/output devices as well as the processors are needed to be considered. Especially, power reduction of disk drives is the most important issue because of the recent extraordinary increase of digital data managed by the systems. This paper aims at the power reduction of the disk array composed of many disk drives, and proposes a new control method of disk drives using query plan of database systems. With showing our own power consumption model of the disk drive, we reveal that significant power reduction can be achievable by the analytical examination based on the model.

♥ 学生会員 東京大学大学院情報理工学系研究科
ueno@tkl.iis.u-tokyo.ac.jp
♦ 正会員 東京大学生産技術研究所
{kgoda,kitsure}@tkl.iis.u-tokyo.ac.jp

1. はじめに

従来，コンピュータシステムにおける消費電力に関しては，主にモバイルコンピュータや組込みコンピュータなどの資源が制限された環境におけるプロセッサの省電力化を中心に検討がなされてきた．近年では，データセンタなどの大規模システムにおいてもその消費電力が問題となりつつあり，また，プロセッサだけでなく周辺の入出力機器を含めたサブシステムの省電力化が求められるようになっている．特に，システムの管理するデータ量が急激に増大している中，ディスクドライブの省電力化は極めて重要な課題である．このため，2000年以降，多くのディスクストレージの省電力化に関する研究が行われている．

ディスクドライブの消費電力の多くは，スピンドルモータとアクチュエータによって消費されていることから，ディスクドライブがアイドルである期間に，ヘッドをアンロードするとともに，ドライブの回転を停止させる(スピンダウンする)ことによって，ディスクドライブの消費電力を削減する方式が一般的である．一定時間ディスクドライブがアイドルである際にスピンダウンするTPM(Traditional Power Management)と称される制御は，既に広く商用ディスクドライブで実装されおり，特に，ラップトップPCやモバイル端末などで利用されている．これは，エンドユーザ環境では，主に対話的アプリケーションが利用されるため，一定のアイドル時間が見込まれる上，システムの応答性能が必ずしも強く求められないため，スピンアップ時のオーバヘッドを比較的許容できることを利用している．

一方，データセンタなどで利用されるディスクストレージの場合，一般に性能要求は高く，スピンアップのオーバヘッドを許容できず，また，単純なTPMによる閾値制御ではディスクをスピンダウンする期間を十分に生み出せない恐れがあり，一般に，TPMをそのまま応用することは困難であるとされている．より高度な制御手法とし，近年，ディスクアクセスの局所性を利用し，データ配置を変更することによりドライブのアイドル期間を生成し，当該ドライブのスピンダウンを計るMAID[2]やPDC[1]と呼ばれる制御や，近年の磁気工学の発展により可能となったディスクドライブの多様な省電力化機能の活用を目指して，スピンダウンされたドライブの復帰時間を短縮する提案[4]や，応答時間に基づきディスクの回転数を調整するDynamic RPM[3]と称される提案が行われているほか，データ配置の調整とディスクドライブ制御を併用するHybernator[5]なる提案もなされるに至っている．

しかし，いずれの手法も，サーバからの入出力アクセスに対して，ディスクストレージは制御情報としては入出力統計情報などのディスクストレージ内で観測可能な情報のみを利用している点において，その制御方式は一概に受動的であると言える．これに対し，本論分では，能動的な制御手法として，データベースシステムの問合せ実行計画に代表される入出力の予定情報をディスクストレージが活用することにより，より高い省電力化効果を達成することを目指す．著者の知る限り，同様の提案は他に見当たらない．

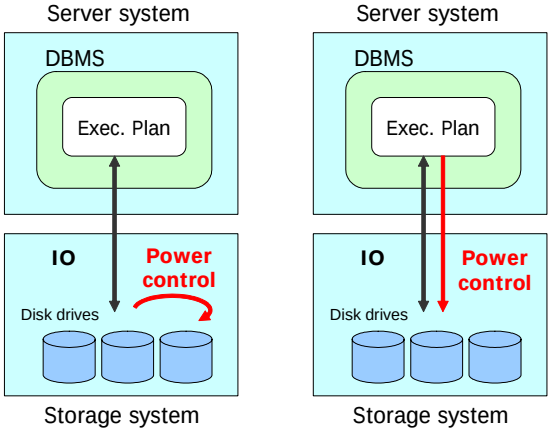
本論文の構成は以下のとおりである．2.では問合せ実行計画を利用した省電力化制御方式を提案する．当該方式の評価のため，3.では実験環境とディスクドライブの消費電力モデルを示し，4.で評価を示す．最後に5.で，論文をまとめる．

2. 問合せ実行計画を利用した省電力化方式

本論分では，ディスクストレージの能動的な制御手法とし

て，データベースシステムの問合せ実行計画を利用した省電力化方式を提案する．図1に，従来方式との比較により，当該方式のアーキテクチャを示す．従来方式では，省電力化制御はディスクストレージの中で閉じて実施されていたのに対し，提案方式では，サーバの有する入出力予定情報に基づきディスクストレージが省電力化制御を行う点に特徴を有する．これにより，従来方式と比較して，高い省電力化効果が期待できる．

本節では，まず典型的なディスクドライブの省電力化機能を説明したのち，提案方式を具体的な事例として多段ハッシュ結合におけるケーススタディを述べる．



(a) Conventional method (b) Proposed method

図1 省電力化方式の比較

Fig.1 Comparison of power reduction methods

2.1 ディスクドライブの省電力化機能

今日のディスクドライブは，多様な省電力化機能を有している．以下では，典型的な省エネルギー型ディスクドライブであるHGST製ATAディスクドライブであるDeskstar T7K250を例に，省電力化機能を説明する．当該ディスクドライブの電力モードは以下の通りである．

- ノーマルアクティブ状態:ディスクドライブが入出力処理を行っている状態．この際，スピンドルが回転し，ヘッドはディスク上に存在する．
- ノーマルアイドル状態:ディスクドライブが入出力は行っていないが，即座に入出力処理を開始可能な状態．この際，スピンドルが回転し，ヘッドはディスク上に存在する．
- アンロード状態:ディスクドライブが入出力処理を開始するのに一定の時間を要する状態であって，スピンドルは回転しているものの，ヘッドはランプに退避している状態．
- 低速回転状態:ディスクドライブが入出力処理を開始するのに一定の時間を要する状態であって，スピンドルは通常より低速で回転しており，ヘッドはランプに退避している状態．
- スタンバイ状態:ディスクドライブが入出力処理を開始するのに一定の時間を要する状態であって，スピンドルは停止しており，ヘッドはランプに退避している状態．
- スタンバイ状態:ディスクドライブへの通電がなされていない状態．この際，ディスクドライブのすべての機構は動作を停止しており，よって，入出力処理を開始するのに一定の時間を要する．

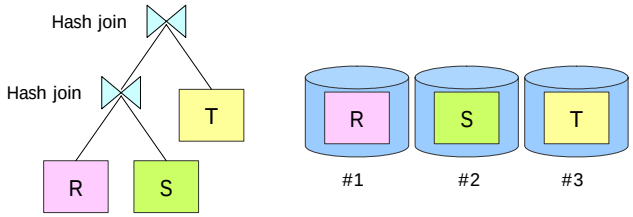
一般に，ノーマルアクティブ状態が最も消費電力が高く，上記の順で，ディスクドライブの部分機構を停止，もしくは減速させることにより，消費電力が低下し，スタンバイ状態では消費電力が0となる．

上記の電力モード間の遷移には，機械的な定常状態間の遷移コストとして時間とエネルギーを要する．たとえば，スタンバイ状態において，入出力要求がディスクドライブに到達した場合，ノーマルアイドル状態へ移行するが，この際には，まず停止しているスピンドルモータを規定の回転速度まで加速させるとともに，引き続いてアクチュエータを動作させヘッドをディスク上に移動させることとなる．これらのモータ制御に係る電力モード制御のオーバヘッドとして無視できない．即ち，ディスクストレージの省電力化制御においては，予測の失敗が性能と消費電力に与える影響は大きいが，サーバの有する入出力予定情報を活用することにより，比較的高い確度で省電力化制御を行うことが可能となることから，著しい省電力化効果の向上が見込まれる．

2.2 多段ハッシュ結合におけるケーススタディ

提案方式を具体的な事例として，図2に示す多段ハッシュ結合におけるケーススタディを述べる．図の通り，関係表R, S, Tがそれぞれディスクボリューム#1, #2, #3に格納されており，((R join S) join T)なるレフトディープ多段ハッシュ結合を行うものとする．当該ハッシュ結合は以下の問合せ実行計画で実施される．

1. 関係表Rを読込み，主記憶上にハッシュ表を作成する．
2. 関係表Sを読込み，主記憶上のハッシュ表を参照し，結合条件に合致したレコードを以って，新たに主記憶上のハッシュ表を作成する．
3. 関係表Tを読込み，主記憶上のハッシュ表を参照し，結合条件に合致したレコードを以って結合演算の結果を出力する．



(a) Execution plan tree (b) Table allocation

図2 Deskstar T7K250の電力モード

Fig.2 Power consumption modes of Deskstar T7K250

この際，フェーズ1を実施中には，ディスクボリューム#1では入出力処理がなされるものの，他のディスクボリュームには入出力がなされない．即ち，上記の問合せ実行計画を活用することにより，明示的にフェーズ1の期間においてはディスクボリューム#2および#3を省電力化することが可能であることがわかる．また，同様の省電力化は，フェーズ2および3においても可能である．これをまとめると，上記の問合せ実行計画に対するディスクストレージにおける省電力化制御は以下の通りとなる．

1. フェーズ1の開始に先立ち，ディスクボリューム#1をノーマルアイドル状態へとし，ディスクボリューム#2および#3を省電力化(例えば，スタンバイ状態へ移動)する．
2. フェーズ2の開始に先立ち，ディスクボリューム#2をノーマルアイドル状態へとし，ディスクボリューム#1

を省電力化する．

3. フェーズ3の開始に先立ち，ディスクボリューム#3をノーマルアイドル状態へとし,ディスクボリューム#2を省電力化する．
4. フェーズ3の終了後，ディスクボリューム#3を省電力化する．

上記の能動的な制御により，従来型の受動制御と比較して，高い省電力化効果が期待される．

3. 実験環境とディスクドライブ消費電力モデルの構築

提案手法の有効性を評価するために，ディスクストレージの消費電力測定環境を構築した．本節では，当該実験環境を述べるとともに，当該実験環境を用いて構築したディスクドライブの消費電力モデルを示す．

図3に実験環境を示す．データベースサーバからディスクドライブへの給電ラインに横河電機製マルチメータWT230を接続することにより，当該ディスクドライブの消費電力を計測する．マルチメータの出力は計測用PCを用いて記録する．また，計測補助用にデータベースサーバへの給電ラインに横河電機製クランプ型電力系CW120を設置し，同様に計測結果を記録する．データベースサーバは，Pentium4 1.5GHzプロセッサと284MB主記憶を有し，OSとしてRedHat Enterprise Linux 3.0 WSを，ディスクドライブとして先述のT7K250を用いる．

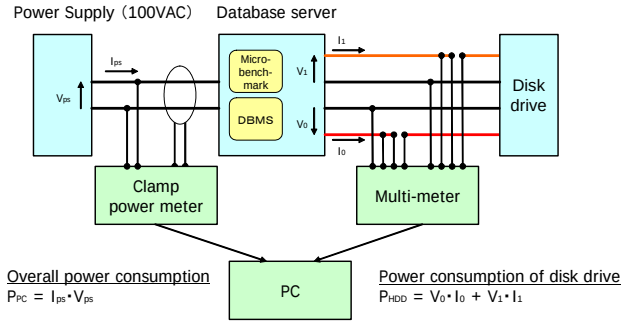


図3 実験環境

Fig.3 Experimental setup

当該ドライブに各種の入出力負荷を与え，その際のスループットと消費電力の関係を解析した．この手順の詳細に関しては，紙面の都合上，論文[6]に譲ることし，本論文では得られた結果をまとめる．当該解析の結果，ディスクドライブが入出力処理を行うノーマルアクティブ状態においては，平均アクセスサイズ別に，以下の関数により，入出力スループットを元に消費電力を解析することが可能となった．

$$P_{4KB} = 0.0849\theta + 5.0176$$

$$P_{16KB} = 0.0701\theta + 5.0324$$

$$P_{64KB} = 0.0686\theta + 4.9859$$

ここに，θは入出力スループット[MB/s]を表し，P_{nKB}は平均アクセスサイズをnKBとした際の消費電力[W]を意味する．また，ノーマルアクティブ状態以外の各電力モードにおける消費電力，および遷移コストは表1の通りであった．なお，ここに，遷移コストは主要な遷移コストを記載し，記載のない遷移は遷移コストが極めて小さい，もしくは実装上遷移が不可能であることを意味する．

表1 消費電力モデルのパラメータ

Fig.1 Power consumption model parameters

(a) Steady-state paramerters

Normal idle	Unloaded	Low rpm	Standby
4.744[W]	3.777[W]	2.223[W]	0.874[W]

(b) Transition costs

Nommal idel → Unloaded	3.5[J], 0.7[s]
Unloaded → Low rpm	15.5[J], 10[s]
Unloaded → Normal idle	3[J], 0.6[s]
Low rpm → Normal idle	37.9[J], 3.2[s]
Standby → Normal idle	83.2[J], 7.2[s]

上記の消費電力モデルを検証するため，データベースサーバにおいて日立製作所製DBMSであるHiRDBと代表的なデータベースベンチマークであるTPC-Hを用い，各問合せの実行に必要とするディスクストレージの消費電力量に関して，マルチメータでの計測とモデルに基づく予測を比較した．結果を図4に示す．この結果，ほとんどの問合せに対して10%未満の誤差で消費電力量を予測することが可能となった．即ち，構築したモデルによって，簡便なスループットの計測に基づき消費電力を高い精度で予測することが可能となった．

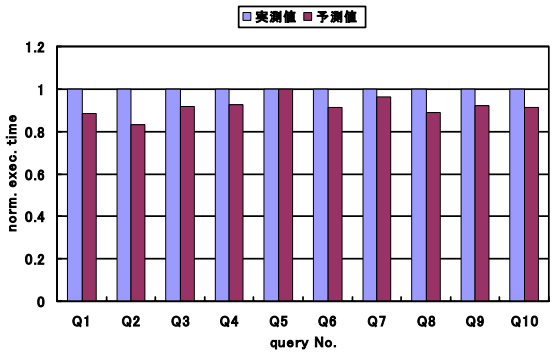


図4 TPC-H問合せによる電力消費モデルの予測誤差

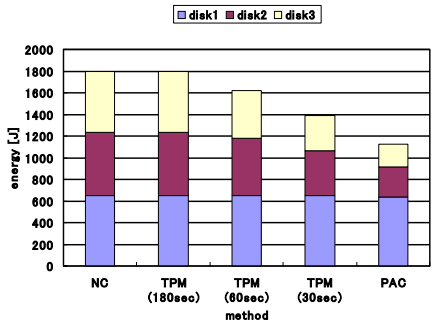
Fig.4 Prediction error of power consumption model using TPC-H queries

4. 問合せ実行計画を利用した省電力化方式の評価

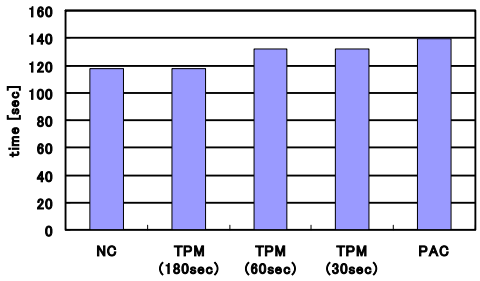
前節の消費出力モデルを用いて，問合せ実行計画を利用した省電力化方式の評価を行った．データセットとしてはTPC-Hを用い，関係表のうち，LINEITEM表をDisk 1に，ORDERS表をDisk 2に，残りの表をDisk 3に格納した．問合せとしては，当該ベンチマークのQ8およびQ9を用い，それぞれについてレフトディープハッシュ結合の実行計画とライトディープハッシュ結合の実行計画により実行し，消費電力量と実行時間を解析した．また，この際，提案手法との比較として，従来方式であるTPMに関して，閾値を180秒，60秒，30秒としたそれぞれのケースについて同様の解析を行った．ディスクドライブの省電力化制御には，スタンバイ状態と低速回転状態を利用し，比較した．

図5に，Q8のレフトディープハッシュ結合に関して，スタンバイ状態を利用した場合の解析結果を示す．なお，NCは省電力化制御を行わない場合，TPMは従来方式としてのTPMによる省電力化制御を行う場合，PACは提案方式による省電力化制御を行う場合をそれぞれ示す．ここに，TPMでは閾値によって省電力化効果が0-20%と変化している一方，PACは能動制御によって35%の高い省電力化効果が達成されている．一方，

実行時間のオーバヘッドは，TPMで0-10%と同様に幅があるものの，PACでは18%であった．



(a) Total energy consumption

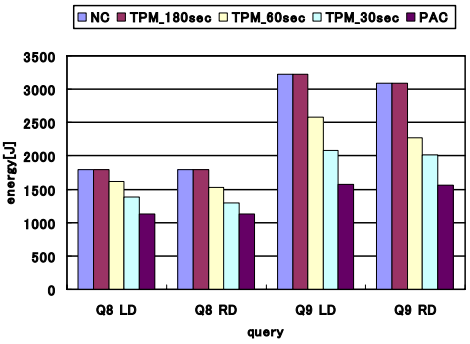


(b) Execution time

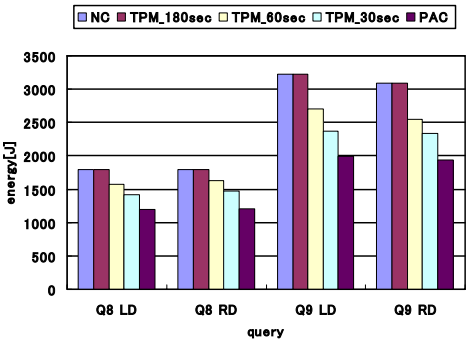
図5 省電力化効果の評価

Fig.5 Evaluation of power reduction effect

図6に，評価実験で行った問合せおよび問合せ実行計画に関して，消費電力量をまとめる．TPM制御では，省電力化効果はおよそ18-34%であるのに対し，PACでは35-50%の省電力化効果を得ることができた．



(a) Result using standby mode



(b) Result using low-rpm mode

図6 省電力化効果の評価

Fig.6 Evaluation of power reduction effect

以上より，従来手法と比較して，提案手法が高い省電力化効果を得られることが明らかになった．

5. まとめ

本論文では，多数のディスクドライブから構成されるディスクアレイの省電力化を目指し，データベースシステムの有する問合せ実行計画を利用した新しいディスクドライブの制御方式を提案した．独自の方式により構築したディスクドライブに関する消費電力モデルを示すとともに，当該モデルに基づき解析的検討により，TPC-Hデータベースベンチマークの問合せ処理において従来方式と比較して高い省電力化効果が得られることを明らかにした．

[文献]

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

[2] D. Colarelli and D. Grunwald. Massive Arrays of Idle Disks for Storage Archive. In Proc. Int'l Conf. on Supercomputing, pp. 1-11, 2002.

[3] 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.

[4] Nexsan Technologies. Disk Based Storage Solutions: The Next Generation Now. Presentation Material, 2005.

[5] 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, pp. 177-190, 2004.

[6] 上野裕也，合田和生，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．電子情報通信学会第 18 回データ工学ワークショップ／第 5 回日本データベース学会年次大会 (DEWS2007), L6-2, 2007.

上野 裕也 Yuya UENO

2005 東京大学工学部電子情報工学科卒業．2007 同大大学院情報理工学系研究科電子情報学専攻修士課程修了．現在，ファナック株式会社勤務．本会学生会員．

合田 和生 Kazuo GODA

2000 東京大学工学部電気工学科卒業，2005 同大学院情報理工学系研究科電子情報学専攻博士課程単位取得満期退学．博士（情報理工学）．現在，東京大学生産技術研究所特任助教．並列データベースシステム，ストレージシステムの研究に従事．本会，情報処理学会，ACM，IEEE CS，USENIX 会員．

喜連川 優 Masaru KITSUREGAWA

1978 東京大学工学部電子工学科卒業．1983 同大大学院工学系研究科情報工学専攻博士課程修了．工学博士．同年同大生産技術研究所講師．現在，同教授．2003 より同所戦略情報融合国際研究センター長．データベース工学，並列処理，Web マイニングに関する研究に従事．現在，本会理事，情報処理学会，電子情報通信学会各フェロー．ACM SIGMOD Japan Chapter Chair，電子情報通信学会データ工学研究専門委員会委員長歴任．VLDB Trustee (1997-2002), IEEE ICDE, PAKDD, WAIM などステアリング委員．IEEE データ工学国際会議 Program Co-chair(99), General Co-chair(05).