Storage Fusion: データベース処理を意識したディスクストレージ省電力化の一考察

合田和生† 喜連川優†

†東京大学 生産技術研究所

概 要

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，当該システムの省エネルギー化は重要な研究課題である．本論文では，上位層のアプリケーションであるデータベースシステムが有する高レベル処理知識を活用してディスクストレージのエネルギー消費量を削減する手法を示す．これまでの省エネルギー化手法が，ストレージシステム内において低レベル制御情報のみを活用していたのと比べ，高い省エネルギー化効果が期待される．

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 [4]，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 [2]．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 [1]．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，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いることから，特にデータインテンシブな IT システムにおいてはストレージシステムによるエネルギー消費はより重要な問題となろう [7]．

スピンドルモータがディスクドライブの主要なエネルギー消費源である．今日のディスクストレージには多数のディスクドライブが具備されていることから，適時にスピンを停止させ，また駆動させることがストレージシステムの省エネルギー化に向けた自然なアプローチである．しかし，スピン制御は機械的操作を伴うため，時間損は数秒から数十秒単位のものであり，これは IT システムの他の構成要素のそれと比較して著しいものである．制御に伴う時間損を克服するべく，如何にスピンを制御するかが，省エネルギー化の鍵となる [6]．

本論文では，ディスクドライブのエネルギー管理に関して，データベースシステムが有する高レベル処理知識を活用してディスクストレージのエネルギー消費量を削減する手法を示す．従来，データベースの問合せ処理は主に高性能を指向していたのに対して，本論文ではエネルギー消費と性能をバランスさせるために問合せ実行計画を活用することを提案する．著者らは論文 [8] 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おいて問合せ処理を遅延させることによる省エネルギー化効果を議論し，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．

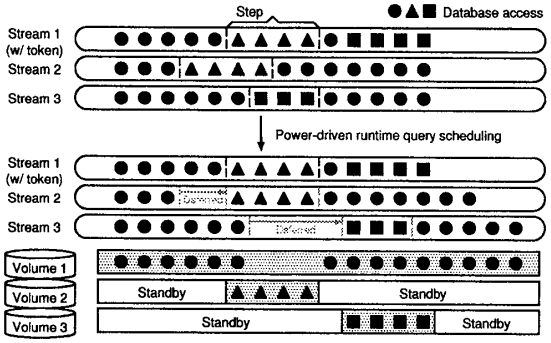

図 1: 遅延化問合せ処理.

2 遅延化問合せ処理

問合せがデータベースサーバに発行されると問合せ実行計画が生成され，当該計画に基づき問合せが処理される．計画では，どのステップでどの関係表や索引をアクセスするかが決められており，当該情報を活用することにより，より効率的なエネルギー管理が期待される．

例えば，2 つの関係表 R と S がそれぞれ，ボリューム V1 と V2 に格納されているデータベースを想定する．R と S を結合する問合せが与えられ，ハッシュ結合が選択されたとする．即ち，まず R を走査することによりハッシュ表を生成し，その後 S を走査することによりハッシュ表の検索を行う．この際，前半のステップではボリューム V1 のみがアクティブであり，後半のステップでは V2 のみがアクティブである．よって，前半のステップでは V1 をスピンアップし V2 をスピンダウンし，後半では逆に V2 をスピンダウンし V1 をスピンアップすることにより，消費エネルギーを削減

Storage Fusion: A Study on Energy Efficient Disk Storage with Assistance of Database Processing
GODA Kazuo† and KITSUREGAWA Masaru†
†Institute of Industrial Science, The University of Tokyo
{kgoda,kitsure}@tkl.iis.u-tokyo.ac.jp

することが可能である．

本論文では，このような能動的なディスクドライブのスピン制御を，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1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．

3 シミュレーション実験

IO トレースに基づくディスクドライブシミュレーション環境と TPC-H ベンチマークを用いて，遅延化問合せ処理によるディスクストレージの省エネルギー化効果を検証した．紙面の制約により，本論文ではその概要を記すに留め，詳細は別稿に譲る．

まず，Linux サーバ上で商用データベースシステムである HiRDB [5] を用いて TPC-H のアドホック問合せを実行し，問合せ実行計画を取得するとともに，カーネルレベルの微視的 IO トレーサを用いて各ステップ毎の IO 処理を取得した．取得した実行計画と IO トレースも用いてシミュレーションを行った．なお，シミュレーションでは，3 つのボリュームからデータベースが構成され，LINEITEM と ORDERS がそれぞれ別箇のボリュームに，残りの関係表が同一ボリュームに格納されるものとした．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．**NC** は一切ディスクドライブのスピン制御を行わないケースであり，これを正規化の基準とした．**FT** は従来型のアイドル時間の閾値に基づくスピンスピンダウン制御 [3] を行うものであり，**PC** は論文 [8] で提案した能動的なスピン制御を行うものである．++が，本論文で提案する遅延化問合せ処理の有効化を意味する．

図2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，能動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

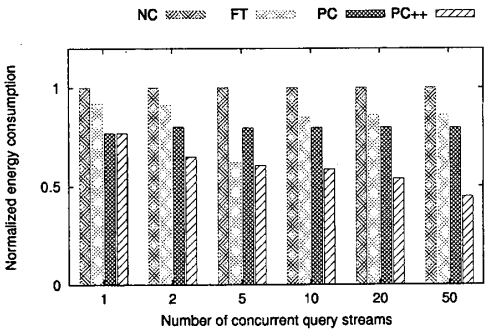

図 2: 省エネルギー効果．(スケールファクタ: 20.0)

優る結果を得た．能動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．

4 おわりに

本論文では，上位層のアプリケーションであるデータベースシステムが有する高レベル処理知識を活用してディスクストレージのエネルギー消費量を削減する手法として，遅延化問合せ処理を示した．また，シミュレーション実験によって，当該手法と能動的なディスクドライブのスピン制御を併用することにより，ディスクストレージの消費エネルギーを約 40-55%削減可能であることが示された．

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．

参考文献

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

[2] B. Rudolph. Storage In an age of Inconvenient Truths. SNW2007Spring, 2007.

[3] 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*, pp. 121–137, 1995.

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

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

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

[7] 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.

[8] 上野, 合田, 喜連川. 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.