社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

コラボラティブグリーンストレージ：データベースシステムとの連携によるディスクストレージ省電力化の構想

合田　和生†　　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,kitsure}@tkl.iis.u-tokyo.ac.jp

あらまし　最近のデータセンタにおいては，ストレージシステムに多数のディスクドライブが組込まれる傾向があり，ストレージシステムは電力消費源として無視できないものとなっている．当該システムの省電力化は，サーバやネットワーク装置のそれと同様に，IT システムにとって重要な課題である．本論文では，**コラボラティブグリーンストレージ**と称するストレージ省電力化の新しいアプローチを提案する．従来の省電力化アプローチでは，サーバ上のアプリケーションと独立にストレージシステムを制御していたのに対し，著者らは，ディスクストレージとアプリケーションとをより緊密に連携させることにより，より著しい省電力化効果を得ることを目指している．省電力型問合せ処理とリモートレプリケーションにおけるケーススタディを紹介し，アプローチの有効性を報告する．
キーワード　グリーンストレージ，データベース，省電力

Collaborative Green Storage: A Vision for Energy Efficient Disk Storage Enabled by Cooperation with Database Systems

Kazuo GODA† and Masaru KITSUREGAWA†

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

Abstract　Recent data centers have been accommodating increasingly larger number of disk drives for storing explosively ever-growing digital data and for improving processing throughput. Electric power consumed by such disk drives have no longer been ignorable, but rather one of primary challenges in IT systems research and development. This paper proposes *collaborative green storage*, a new approach for energy saving of disk storage. In contrast to conventional approaches, which tried to control power modes independently of server applications, our main idea is in the collaboration between disk storage and server applications for obtaining significant energy saving. Two case studies, power aware query processing and green replication, would validate the potential benefits of the approaches.
Key words　Green Storage, Database, Power Saving

1. ま え が き

本論文ではデータセンタにおける，殊にエンタープライズシステムにおけるディスクドライブの消費電力の削減方式について議論する．当該システムでは，爆発的に増大するデータを確実に管理し，また，高速に処理するために，極めて多数のディスクドライブが組込まれる傾向にある．例えば，典型的なエンタープライズアプリケーションと言えるオンライントランザクション処理については，TPC-C [1] が業界標準のベンチマークであり，IBM のシステム [2] が最近までトップレコードを保持していたが，そのシステム構成は 1 万台以上のディスクドライブが 1 台のサーバに接続されたものであった．1 万台は性能を追求した極端な例であるが，現にエントリーレベルのシステムにおいても数十台の RAID 構成 [3] は珍しいものではなく，最近の高性能サーバの能力を最大限に発揮しようとした場合，相当数のディスクドライブを並列駆動する必要があることは明らかである．自ずと，システム全体の中で，これらのディスクドライブによって消費される電力は無視できないばかりか，むしろサーバのそれを凌ぐ場合もある．同じ TPC-C ベンチマーク委員会に投稿された各社システムの消費電力について Oracle が調査したところ，システム全体の消費電力のうち，約 8 割がディスクドライブによって消費されていた [4]．エンタープライズシステムの省電力化を目指す場合に，ディスクドライブを中心とするディスクストレージの省電力化は極めて重要な課題

であろう．

ディスクドライブの消費電力はそのほとんどがスピンドルモータによっている．当該モータはデータが記録される金属円盤を空気抵抗に逆らって回転させるものであるが，ディスクアクセス遅延を短縮するために回転は 15000rpm へと高速化しており，これによる電力消費が著しい [5]．よって，ディスクドライブの消費電力を削減するためには，ディスクアクセスのある期間のみ当該モータを駆動し，ディスクアクセスのない期間に当該モータを停止させることが基本である．ところが，他の計算機部品と異なり，スピンドルモータは機械駆動部品であり，特にスピン開始に伴う時間損とエネルギー損はそれぞれ数十秒と数百ジュールにのぼる [6]．スピンドルモータの停止によって消費電力を削減するためには，多くの場合，数十秒以上に渡って全くディスクアクセスのないアイドル時間が必要であり，しかもオンラインの制御系がこれを的確に予測してスピンドルモータの停止と再開を指令する必要がある．

エンタープライズシステムについては，図 1(a) に示すように，これまでに現行のストレージインターフェースの上でどれほどの省電力化の効果が期待できるかという観点で主にディスクストレージの制御系に関する研究が進められてきた．すなわち，サーバにおける制御とディスクストレージにおける制御は各々互いに独立に行われてきた．TPM と称される制御方式 [7–9] では，ディスクアクセスが一定時間以上行われていない場合にその後も相当時間，アクセスが行われない可能性が高いとの経験則を利用して，閾値ベースのアイドル時間に基づきディスクドライブのスピンダウンを行うものである．アーカイブ目的のストレージシステムでは良好に機能することが分かっている．最近のディスクストレージには MAID [10] と呼ばれる機能が実装されている．これは，まさに TPM の実装に他ならないが，実は当初 MAID が提案された論文 [11,12] では，閾値ベースのスピンダウン制御に加えて，アクセス局所性を活用したデータマイグレーション機能が提案されている．ディスクストレージ中の特定のディスクドライブにホットなブロックをコピーし，当該ディスクドライブをキャッシュとして利用し，他のディスクドライブへのアクセスをなるべく減らすことにより，当該ディスクドライブのアイドル時間を捻出しようとするものであり，オンラインとアーカイブの中間に置くニアラインと呼ばれるストレージにおいても有効に機能する場合がある．上記に代表される既存のアプローチはいずれも一定の消費電力の削減効果が認められているものの，なおも削減の余地が大きい．

これらに対して，本論文では既存のストレージインターフェースに捕われず，むしろ新しいストレージとサーバの接続関係が可能となった場合に，どれほどの省電力化の効果が期待できるかという視点に立ち，コラボラティブグリーンストレージと称する新しい消費電力削減のアプローチを提案する．現行のディスクストレージは基本的には READ と WRITE という 2 つの低レベルアクセス命令のみでサーバと通信を行っており，自ずと，消費電力の削減を行う制御系はディスクストレージ内に閉じる．インターオペーラビリティの観点からは優れたアーキテクチャではあるものの，もはや最近のディスクストレージは単

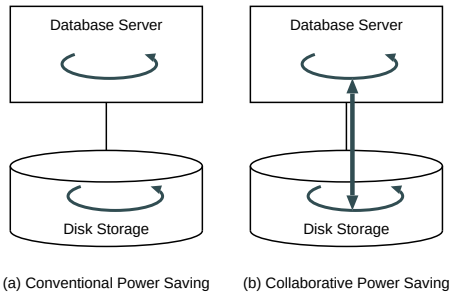

図 1　ディスクストレージの省電力化手法の比較.
Fig. 1　Comparison of power saving methods for disk storage.

なる受動的な記憶装置ではなく，高度なプロセッサを具備して当該資源を活用して様々な管理アプリケーションが動作する装置に成熟している．図 1(b) に示す通り，より高いレベルのインターフェースを規定し，サーバ上のアプリケーションと高度な連携をはかることにより，サーバ上のアプリケーションまでを含んでシステム全体を見渡したグローバルな制御が可能となり，より一層の消費電力の削減効果が得られる可能性がある．

本論文では，コラボラティブグリーンストレージに関して著者らが行ってきている 2 つのケーススタディを報告する．2. で示す省電力型問合せ処理は，意思決定支援システムなどに見られる比較的長時間を要する問合せ処理に関して，データベースサーバが問合せ処理に先立ち生成する問合せ実行計画なる実行情報をディスクストレージに開示するとともに，データベースサーバがディスクストレージの消費電力を意識したスケジューリングを行うことにより，消費電力の削減を目指すアイデアである．また，3. で示すグリーンレプリケーションは，所謂ディザスタリカバリ等に用いられるレプリケーションにおいて，遠隔データセンタにおけるディスクストレージへの書込みをデータベースの更新ログの情報を活用して高度にスケジュールすることにより，遠隔データセンタのディスクストレージの消費電力の大幅な削減を目指すものである．おわりに 4. でケーススタディと今後の検討課題を纏める．

2. 省電力型問合せ処理

本節では，特に意思決定支援システムなどにおけるデータベースサーバとディスクストレージが相互に連携することにより，問合せ処理に掛かる消費電力を削減するアイデアとして，プロアクティブ電力制御と問合せ処理の遅延化可能性による実行時スケジューリングを示す．

2.1 プロアクティブ電力制御

一般に，データベースサーバに問合せが与えられると，当該サーバはまず問合せ実行計画なる実行情報を生成し，当該実行情報に従い問合せ処理を行う．当該実行情報を同時にディスクストレージに開示することにより，ディスクストレージが高い確度で近い将来のディスクアクセスを予測し，能動的にディスクドライブの電力制御を行うことが可能となり，よって高効率の電力制御を実現することが期待される．

関係データベースにおいて R と S という 2 つのテーブルを

ハッシュ結合なるアルゴリズムで結合する例を示す．この場合，まずデータベースサーバは R が格納されたディスクドライブから R を読出して，主記憶上のバッファに結合キーによるハッシュテーブルを生成し，レコードをハッシュテーブルに格納する．その後，S が格納されたディスクドライブから S を読出し，レコード毎にハッシュテーブルの検索を行い，結合条件に合致したレコードを出力する．問合せ実行計画には当該手続きが記録されており，当該情報をディスクストレージに開示することにより，R が格納されたディスクドライブと S が格納されたディスクドライブをいつどのようにアクセスするかをディスクストレージは概ね予測することが可能となる．よって，R と S がそれぞれ異なるディスクドライブに格納されている場合，R をアクセスしている期間には S が格納されたディスクドライブをスピンダウンし，S のアクセスが始まる事前に当該ディスクドライブをスピンアップしておき，さらに，S のアクセスが始まったあとには R が格納されたディスクドライブをスピンダウンすることができるだろう．これは，旧来型の受動的な電力制御と比べて問合せ処理の性能に掛かる副作用が少なく，また，高い電力削減効果があり，有益性が高い．

著者らは [13] において，TPC-H [14] なる意思決定支援系ベンチマークに対するシミュレーション実験を行ったところ，旧来型のアイドル時間閾値に基づく電力制御では，問合せ処理に掛かる電力量の削減効果が最大でも 15-35%に留まる一方，閾値設定によっては著しい性能オーバヘッドを観測した．これに対して，プロアクティブ電力制御を行ったところ，多くの場合で 35-50%の消費電力の削減効果を得たほか，性能オーバヘッドは殆ど観測されず，当該方式が高い潜在的優位性を有することを確認した．

2.2 問合せ処理の遅延化可能性による実行時スケジューリング

2.1 では，単一の問合せに関してその実行情報をディスクストレージが活用して電力制御を行うことの有益性を示した．本節では，更にこれを複数の問合せが同時に処理される場合に拡張したい．

複数の問合せが同時に処理される場合，同様にプロアクティブ電力制御を行うことが可能である．しかし，各々の問合せが独立の処理されている場合，ある問合せにおいては特定のディスクドライブがアクセスされないのでこれをスピンダウンして消費電力を削減しようとするのと同時に，別の問合せが同じディスクドライブをアクセスしてしまうと，総じての消費電力の削減効果が低くなる可能性がある．本節では，同時に処理される複数の問合せの処理を相互に調停することにより，電力削減効果を高める遅延化可能性に基づく実行時スケジューリングを示す．

図 2 に基本アイデアを図示する．データベースサーバに複数の問合せ系列が到達しているとしたとき，特定の系列のみがディスクドライブのスピンアップを行うことが可能であり，それ以外の系列が問合せ処理のためにスピンダウン中のディスクドライブをアクセスしようとする場合には，原則的に他の系列によって当該ディスクドライブがスピンアップされるまで待つ

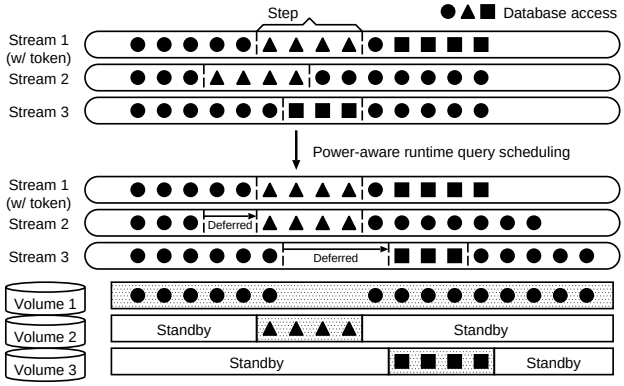

図 2　遅延化可能性による実行時スケジューリング．
Fig. 2　Runtime scheduling based on deferability of queries.

ものとする．系列間の公平性のための制御系を用意する必要が生じるが，同時に処理される問合せ間でディスクドライブの電力状態を意識した相互スケジューリングを行うことにより，総じて消費電力の削減が期待される．

著者らは [15] において，同じく TPC-H を対象としたシミュレーションを実施したところ，単にプロアクティブ電力制御を行うだけのものに関しては電力削減効果が概ね 20%程度に留まったのに対して，スケジューリングを併用することにより，40-55%程度の省電力化を達成可能であるとの結果を得た．

従来，エンタープライズシステムでは，ディスクドライブの電力制御はディスクストレージ内に閉じて行われていたが，これと比較して，データベースサーバとディスクストレージを連携させる新しいアプローチが著しい優位性を有する見込みを得ている．

3. グリーンレプリケーション

本節では，所謂ディザスタリカバリ等に用いられるレプリケーションにおいて，遠隔データセンタにおけるディスクストレージの消費電力を削減するアイデアを示す．

エンタープライズシステムでは，一次系で動作するオンライントランザクション処理などのデータを保護するために，二次系にレプリカを設け，万が一のシステム障害や災害時に迅速にサービスを再開できるように備えることが求められていることが多い．このとき，図 3 に示すように，一次系から二次系に更新ログを転送する必要があり，二次系では受信した更新ログをデータベースに反映する必要がある(注1)．

著者らの基本アイデアは，二次系に転送された更新ログを一時的にログボリュームに格納し，データベースへの反映を遅延し，一定量，未反映の更新ログが溜まってから一括して更新ログを反映するものである．これにより，二次系においてはデータベースの格納されているディスクドライブへのアクセスを無くし，よって当該ディスクドライブをプロアクティブにスピンダウンすることができる．万が一，一次系においてサービスが

(注1)：ここでは簡単のためにデータベースの更新ログを二次系に転送するレプリケーションのみを説明しているが，同じアイデアは他のブロックレベルのレプリケーション等にも適用可能である．

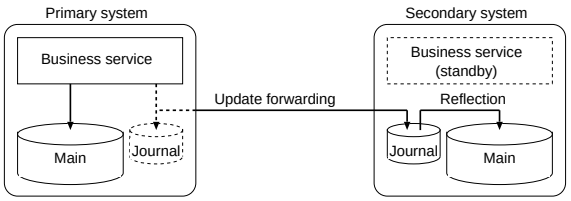

図 3　業務継続を目的としたレプリケーションシステム．
Fig. 3　A replication system for business continuity.

停止した場合に，二次系においてサービスを再開しなければならず，この際，サービスを再開するためには未反映の更新ログを全て反映する必要がある．一般に，サービス再開のために許容されうる時間損は RTO (Recovery Time Objective) と呼ばれシステムによって異なるが，RTO 制約内でサービスを再開することがエンタープライズシステムにとっては必須であることから，二次系において更新ログの反映をどれほど遅延化できるかは，一括反映のスループットに依存する．すなわち，二次系に到達する更新ログのレートに対して，一括反映のレートが高いほど，より長時間，ディスクドライブをスピンダウンすることが可能である．一括反映のスループットを格段に高める技法として著者らは更新ログのコンパクションなる手法 [16,17] を著者らは提案している．なお，通常，データベースシステムにおけるストレージ構成では，ログボリュームに用いられるディスクドライブの数は，データベースに用いられる数と比べて相当に小さいため，上記の手法による省電力化効果は高いと言えよう．

著者らは，論文 [16,17] において，TPC-C ベンチマークに対してハイエンドディスクドライブならびにミッドレンジディスクライブを模擬したシミュレーションを実施した．旧来型の同期レプリケーションを用いる場合，二次系のディスクストレージは一切，電力制御を行うことが出来ないのに対して，提案手法を用いることにより，100 秒程度の RTO 追加時間を認めると，二次系のディスクストレージについて 80-85%の高い消費電力の削減効果が得られる結果を得た．

4. むすび

本論文ではコラボラティブグリーンストレージと称する新しい消費電力削減のアプローチについて，省電力型問合せ処理とグリーンレプリケーションなる 2 つのケーススタディを示した．それぞれ，従来手法と比較して高い優位性を有するものの，データベースサーバとディスクストレージのインターフェースの変革を伴う．現時点においては，アプローチの潜在的な優位性をケーススタディ毎に確認しているに過ぎないが，例えば，SNIA なる業界標準団体は SMI-S [18] なるストレージ管理プロトコルの標準化に取り組んでおり，今後は新たなインターフェースのあり方に関して検討を深める必要があるだろう．

謝辞

本研究の一部は，文部科学省次世代 IT 基盤構築のための研究開発「非順序型実行原理に基づく超高性能データベースエンジンの開発」の助成により行われた．

文　献

[1] Transaction Processing Performance Council, “TPC-C, an online transaction processing benchmark,” http://www.tpc.org/tpcc/.
[2] Transaction Processing Performance Council, “TPC-C Result Highlights: IBM Power 595 Server Model 9119-FHA,” http://www.tpc.org/tpcc/results/tpcc_result_detail.asp?id=108061001.
[3] D. Patterson, G. Gibson, and R. Katz, “A Case for Redundant Arrays of Inexpensive Disks,” Proc. ACM SIGMOD Conf., pp.109–116, 1988.
[4] M. Poess and R.O. Nambiar, “Energy cost, the key challenge of today’s data centers: a power consumption analysis of TPC-C results,” Proc. Int’l. Conf. on Very Large Data Base, pp.1229–1240, 2008.
[5] D. Anderson and W. Whittington, “Hard Drives: Today & Tomorrow,” Tutorial, USENIX Conf. on File and Storage Tech., 2007.
[6] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, “Reducing Disk Power Consumption in Servers with DRPM,” IEEE Computer, vol.36, no.12, pp.59–66, 2003.
[7] F. Douglis, P. Krishnan, and B. Bershad, “Adaptive disk spin-down policies for mobile computers,” Proc. USENIX Symp. on Mobile and Location-Independent Computing, pp.121–137, 1995.
[8] P. Krishnan, P.M. Long, and J.S. Vitter, “Adaptive disk spindown via optimal rent-to-buy in probabilistic environments,” Int’l Conf. on Machine Learning, pp.233–330, 1995.
[9] R.A. Golding, P. Bosch, C. Staelin, T. Sullivan, and J.Wilkes, “Idleness is not sloth,” Proc. USENIX Tech. Conf., pp.201–212, 1995.
[10] F. Moore and A. Guha, “Introducing COPAN Systems MAID Architecture (Massive Array of Idle Disks),” White Paper, Copan Systems, 2004.
[11] D. Colarelli, D. Grunwald, and M. Neufeld, “The Case for Massic Arrays of Idle Disks (MAID),” Proc. USENIX Conf. on File and Storage Tech., 2002.
[12] D. Colarelli and D. Grunwald, “Massive Arrays of Idle Disks for Storage Archive,” Proc. ACM/IEEE Conf. on Supercomputing, pp.1–11, 2002.
[13] 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,” 日本データベース学会 Letters，vol.6，no.1，pp.85–88，2007.
[14] Transaction Processing Performance Council, “TPC-H, an ad-doc, decision support benchmark,” http://www.tpc.org/tpch/.
[15] 合田和生， Wenyu Qu，喜連川優，“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,” 電子情報通信学会 データ工学ワークショップ，pp.D5–2，2008.
[16] 合田和生，喜連川優，“ログ転送を用いたディザスタリカバリシステムにおけるディスクストレージの省電力化方式の検討,” 日本データベース学会 Letters，vol.6，no.1，pp.69–72，2007.
[17] K. Goda and M. Kitsuregawa, “Power-aware Remote Replication for Enterprise-level Disaster Recovery Systems,” Proc. USENIX Tech. Conf., pp.255–260, 2008.
[18] Storage Network Industry Association, “A Dictionary of Storage Networking Terminology,” http://www.snia.org/education/dictionary/.