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

A Study on Power Reduction Method of Disk Storage for Log Forwarding Based Disaster Recovery Systems

合田 和生♥　　喜連川 優♥

Kazuo GODA　Masaru KITSUREGAWA

本論文では，ログ転送方式による遠隔コピーを利用したディザスタリカバリシステムに関して，ログ転送によりサイト間でデータ一貫性が保持される点に着目した省電力化方式を提案するとともに，ベンチマークを用いた評価実験によって，100 [s]程度の RTO 制約の下で，二次サイトにおけるディスクストレージの 85%程度の省電力化が達成されることを示し，ディザスタリカバリシステムの運用コスト低減に大きく寄与することを示す．

In this paper, a new power reduction method is proposed focusing on a disaster recovery system based on log forwarding remote copy. The proposal exploits the site-level transactional recoverability elegantly brought by the log forwarding. For validation, basic evaluations using benchmark are disclosed, showing that 85% energy saving of secondary-site disk storage can be achieved under the RTO constraint of 100 [s]. The contribution opens the gate for significantly reducing operational cost of disaster recovery systems.

1. はじめに

テロやハリケーンなどの予測不可能な災害による業務の停止は，社会や国家に甚大な影響を与えることが明らかになっている．テロ等による金融機関の業務停止は，一社あたり毎時約645万ドルの損失を生むと推測されており[4]，また，米国で2005年に発生したハリケーンKatrinaでは，実際に1000億ドル以上の損害が発生し，うち損害保険の支払い対象が34億ドルにのぼった[7]．予測不可能な災害時における業務継続を目的とした具体的な対策は喫緊の課題であり，米国SEC Rule17a[9]や英国BS25999[1]に代表される国家レベルの法規制が導入されるようになりつつある．

一方，近年の業務は高度に情報化しており，コンピュータシステムによる業務継続のためには，地理的遠隔地にバックアップ用の二次サイトを設けるディザスタリカバリシステム[6,13]を構築することが必須要件である．しかし，当該システムは物理的にサイトレベルの冗長性を課すため，必然的に運用コストは倍化する．即ち，二次サイトの運用コストの低減が極めて重要な課題である．

従来，サーバやディスクアレイなどのハイエンドサブシステムの運用に関しては，人的管理コストが主に重要視されてきたが，近年では，その消費電力が無視できなくなっており，とりわけ，データ量が著しく増大している今日では，ストレージシステムの消費電力の削減が新たな課題として注目されている．本論文では，論文[12]に基づき，ログ転送方式による遠隔コピーを利用したディザスタリカバリシステムに関して，ログ転送によりサイト間でデータ一貫性が保持される点に着目した省電力化方式を提案する．また，ベンチマークを用いた評価実験によって，二次サイトの省電力化と業務継続品質の確保の双方が達成されることを明らかにし，ディザスタリカバリシステムの運用コスト低減に大きく寄与することを示す．なお，著者らの知る限り，同様の研究開発は他に見あたらない．

本論文の構成は以下の通りである．2. においては，ディザスタリカバリシステムと遠隔コピー技術をまとめる．3. においては，同期ログ転送方式の特徴に着目した二次サイトのディスクストレージ省電力化手法を提案し，4. においては，3. の議論を踏まえ，ベンチマークを用いた評価実験を示す．5. においては，関連研究をまとめ，最後に6. において，本論文をまとめるとともに，今後の課題を示す．

2. ディザスタリカバリシステムと遠隔コピー技術

一般に，コンピュータシステムにおけるディザスタリカバリシステムとは，主たる一次サイトに対して，通信路で接続された地理的遠隔地にバックアップ用の二次サイトを設けることにより，災害発生時に二次サイトで業務を継続するものである．ディザスタリカバリシステムでは，平時においては，一次サイトにおいて業務を実施し，更新されたデータを二次サイトへ遠隔コピーを用いて転送するとともに，災害発生によって一次サイトが利用できない際には，一次サイトから系の切替えを行い，二次サイトにおいて業務を継続する．ディザスタリカバリシステムにおける業務継続の品質を規定することを目的として，RTO (Recovery Time Objective)とRPO (Recovery Point Objective)なる２つの指標が広く用いられる．RTOは，災害発生によって一次サイトが停止した後，二次サイトにて業務を継続するまでの時間を意味する．一方，RPOは，災害時に二次サイトにおいて業務を継続する際に，過去のどの時点のデータを以って復旧可能であるかを表す．即ち，RTOは災害時のサービス停止時間を，RPOは災害時のデータ損失可能性をそれぞれ意味することから，両者は共に小さい値であることが望ましい．

ディザスタリカバリシステムにおいては，平時における一次サイトから二次サイトへの遠隔コピーが不可欠な機能である．従来より，遠隔コピー技術としては，同期転送(Synchronous Forwarding)並びに非同期転送(Asynchronous Forwarding)なる2つの基本的な方式が用いられている．前者は，災害時に二次サイトで業務を継続した際に失われるデータが無いことを保証することができる一方，特に広域災害等を念頭としたディザスタリカバリシステムにおいては，一般にサイト間の距離が長く，一定の通信遅延があることから，一次サイトの入出力性能へ無視できない副作用を有する．対して後者は，一次サイトの入出力性能は，サイト間の距離に影響を受けない一方，災害時に二次サイトで業務を継続した際に失われる可能性を排除できない．

♥ 正会員　東京大学生産技術研究所
{kgoda,kitsure}@tkl.iis.u-tokyo.ac.jp

同期転送並びに非同期転送の双方の問題を解決するために，データベースシステムの記憶管理の特徴を利用した同期ログ転送(Synchronous Log Forwarding: SLF)なる新しい方式が提案されている[14]．当該方式では，データベースシステムの管理する記憶空間が，データベース本体を格納するデータボリュームと，その更新記録であるログを格納するログボリュームから構成されることに着目し，ログボリュームのみを遠隔地へ同期転送し，遠隔地においてログボリュームに転送されたログをデータボリュームに適用する．即ち，ログは同期転送されることから，遠隔地の二次サイトのデータベースシステムは，論理的に一次サイトのデータベースシステムに対して常に最新のデータベースを有する一方，データボリュームを転送しないことから，サイト間でのデータ通信量を削減し，また，同期点を抑制することが可能である．特にハリケーンや地震などの広域災害を想定した場合，ディザスタリカバリシステムにおけるサイト間距離は長くなる傾向があるが，当該方式では，同期方式と同様に，サイト間でトランザクション一貫性を保持するとともに，非同期方式と同程度に，一次サイトのデータベースシステムがサイト間距離によらず性能を導きだすことができる点に特徴がある．

3. 同期ログ転送におけるディスクストレージ省電力化

3.1 基本アイデアと省電力化の見積り

同期ログ転送に関しては，ログは同期方式で転送されることから，災害によって一次サイトが停止したとしても論理的に失われるデータは存在せず，RPOがゼロであることを保証できる．一方，転送されたログは二次サイトで必ずしも常時適用する必要はなく，その頻度は制御可能である．本論文では，同期ログ転送におけるログ適用の非同期性を利用し，ログ適用が行われない一定期間に関しては，データボリュームを構成するディスクドライブの省電力化制御を行い，総じて二次サイトのディスクストレージ全体の省電力化を行う．なお，この際，RTOと節約電力の間には相互依存の関係が成り立つが，以下にモデルを用いた解析を行い，その効果を定量的に議論する．

同期転送されたログに関しては，二次サイトにおいて，一定の周期を以って一定のバッチ周期を以って非同期的に適用されるとする．システム制約として最大RTOをT_{RTO}とした場合，ディスクストレージの消費電力を最も削減可能なバッチ周期T_{wnd}およびバッチ周期内のログ適用期間T_{apl}は，以下のとおりとなる．

$$T_{wnd} = \frac{R_{apl}^2}{(R_{apl} - R_{gen}) \cdot R_{gen}} \left(T_{RTO} - \max_t T_{up}(t)\right)$$

$$T_{apl} = \frac{R_{apl}}{R_{apl} - R_{gen}} \left(T_{RTO} - \max_t T_{up}(t)\right)$$

ここに，R_{gen}は一次サイトにおいて生成される単位時間あたりのログレコード数，R_{apl}は二次サイトにおいて適用可能な単位時間当たりのログレコード数，$T_{up}(t)$は時刻tにおいてディスクストレージを即座にアクセス可能な状態へ遷移させるのに必要な時間を意味する．

この際，図1に示す最も代表的な3つの消費電力モードを有するディスクドライブを対象に議論の焦点を絞ると，データボリュームに関しての，ディスクドライブあたりの同期ログ転送時の平均的な消費電力P_{SLF}は以下の通りとなる．

$$P_{SLF} = \frac{T_{apl}P_0 + (T_{wnd} - T_{12} - T_{apl} - T_{21})P_2 + E_{12} + E_{21}}{T_{wnd}}$$

ここに，P_iは消費電力モードiにおける定常的な消費電力，Eijおよび$T_{i,j}$は消費電力モードiからjの遷移に必要な電力量及び時間を意味する．なお，一般にほとんどのディスクドライブでは，モード0とモード1との間の遷移電力量，並びに遷移時間は0と見なすことができるため，本論文でも同様に扱うものとする．なお，上記では簡単のために，3つの消費電力モードに限定した議論を行ったが，同様に議論はより多くの消費電力モードを有するディスクドライブにも容易に活用することが可能である．

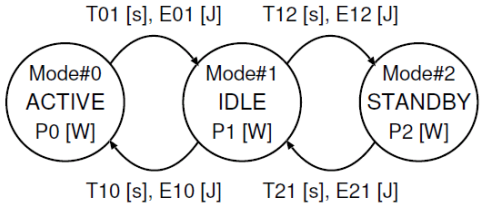

図1 典型的な3モードディスクドライブの消費電力の状態遷移モデル

Fig.1 State transition model for power consumption of typical three-mode disk drive.

3.2 有効性に関する考察

同期ログ転送による省電力化効果に関しては，ログ生成レートとログ適用レートの比率が寄与する割合が高い．即ち，R_{apl}/R_{gen}が大きい程，二次サイトでデータボリュームを省電力化のために制御する余裕があるため，その効果が顕著になる．このため，省電力化効果を得るためには，二次サイトにおけるログ適用の高度化が不可欠である．本論文では，ディザスタリカバリシステムの省電力化に焦点を当てているため，高度化に関する詳細な議論は別稿に譲り，その概要を説明する．一般に，データベースのログ適用においては，蓄積されたログレコードを，ログレコードが記録された順序で適用するが，この際のデータボリュームへのアクセスはランダム書き込みとなることが多く，ログ適用の性能向上に障壁となる．著者らは，論文[11]において，ログ適用の高速化に関して，ログ畳み込み(log folding)とログ整列(log sorting)なる独自の処理方式を提案している．即ち，蓄積されたログに関して，一定のログレコードを主記憶上のログウィンドウバッファに読み込み，同一レコードに関する複数の更新を意味するログレコードを集約するとともに，適用時にログレコードをその適用先のディスクドライブ上のアドレスにて並び替える．これにより，適用ログ数を削減すること，並びにログ適用時の入出力の連続性を高めることが可能となり，ログ適用の大幅な高速化が期待される．TPC-Cベンチマークを用いた著者らの実験では，当該方式により，概ね20倍から50倍の高速化を達成することが可能であることが分かっている．また，他の高度化手法の提案として論文[15]などがあり，同様に著しい高速化が期待される．

なお，上記で示したディザスタリカバリシステムの二次サイトにおけるディスクストレージの省電力化は，データボリュームを構成するディスクドライブの消費電力モードを制御する一方，ログボリュームに関しては操作しない．データボリュームに対する省電力化効果が，ディスクストレージ全体の省電力化にどの程度寄与するかを議論する必要がある．一例として，2006年12月11日現在のTPC-Cベンチマークにおいて公表された性能上位5件のシステムに関して，その構成

を検証すると，4件においてストレージシステムを構成するディスクドライブの95%以上がデータボリュームに利用されており，残りの1件においても90%以上が同様であり，ログボリュームに利用されるディスクドライブは極めて少量であることがわかる．このことから，高い性能を指向するトランザクション処理システムにおいては，ストレージシステムの物理資源の大部分は，データボリュームに利用されており，このため，データボリュームの消費電力を大幅に削減することは，ディスクストレージ全体の省電力化に極めて有効な手段であることが分かる．

4. 省電力化効果の評価

本節では，提案手法によるディスクストレージの省電力化効果の評価として，TPC-Cベンチマークを用いて，ログ適用の高速化技法によるディザスタリカバリシステムの二次サイトのディスクストレージ全体の省電力化効果を検証する．ストレージシステムの構成としては，先述のTPC-Cベンチマークで公表された順位1位のシステムを参考にモデル化を行った．ただし，ディスクドライブの電力消費モデルとしては図1に示す3つの消費電力モードを有するIBM Ultrastar 36ZXを利用した．モデルパラメータを表1に示す．ログ適用の高速化としては，著者らが論文[11]で示した手法を商用DBMSであるHiRDBを用いて簡易実装し，ウェアハウス数を10としたTPC-Cベンチマークスキーマに対して，10万トランザクションを生成し，得られた約173万個のログレコード，並びにその適用トレースを用いて実験を行い，省電力化効果を解析した．

表1 消費電力モデルパラメータ（IBM Ultrastar 36ZX）
Tab.1 Parameters of power consumption model (IBM Ultrastar 36ZX).

Steady-state power consumption		
P_0	P_1	P_2
39 [W]	22.3 [W]	4.15 [W]

Transition delay and energy			
T_{01}	T_{10}	T_{12}	T_{21}
0[s]	0[s]	15[s]	26[s]
E_{01}	E_{10}	E_{12}	E_{21}
0[J]	0[J]	62.25[J]	904.8[J]

図2に検証結果として，ログウィンドウバッファ長，並びにRTO双方の制約に対する省電力化効果をまとめる．本実験では，2MBのログウィンドウバッファに対しては，十分なログ適用の高速化を行うことができず，相対的に消費電力モード制御のオーバヘッドが大きくなるため，ディスクストレージ全体の消費電力を増大させる結果になった．一方，8MB以上のログウィンドウバッファを用いた場合，約30%から85%程度の消費電力を削減することが可能となった．また，RTOの変化に対しては，10[s]以下のRTO制約の下では，ディスクドライブの制御遅延のために省電力化制御そのものを実施することができなかったが，100[s]以上のRTO制約の下では，グラフに示す通り，制御が可能であり，消費電力を削減することができた．前節での見積りと同様に，100[s]以上においては，RTO制約が省電力化効果に与える影響は限定的であることが分かった．

以上より，TPC-Cベンチマークを用いた検証によって，提案手法が有効に機能することが明らかになった．

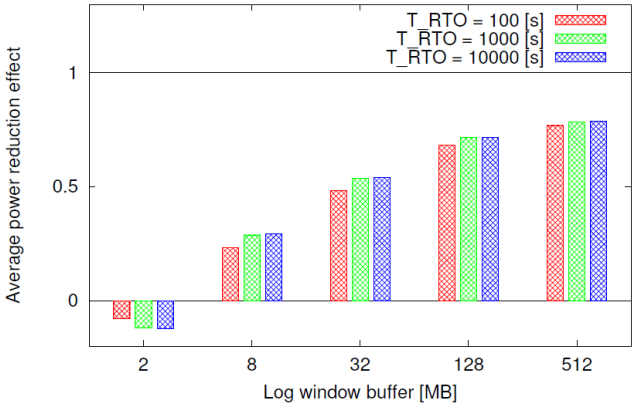

図2 ディスクストレージの省電力化効果（TPC-C）
Fig.2 Power redunction effect on disk storage (TPC-C).

5. 関連研究

ストレージシステムの省電力化に関する研究は，特に2000年以降に活発に行われるようになっている．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ており，また，ラップトップPCやモバイル端末などで利用されている．

ストレージシステムの省電力化に関する代表的な研究としては，主にディスクアクセスの局所性を活用する研究が行われてきた．MAID (Massive Array of Idle Disks)[3]は，ディスクストレージ内部の記憶空間をホット領域とコールド領域に分割し，ホット領域をコールド領域のキャッシュ空間として利用する．データアクセスの局所性を利用して，コールド領域のディスクドライブをアイドル化することにより，当該ドライブを省電力化させることを目指す．また，PDC (Popular Dara Concentration)[2]はMAIDと同様にディスクストレージを複数の領域に分割するものであるが，領域を他領域のキャッシュとして利用するのではなく，データ移送を行う点が異なる．また，単にディスクドライブの回転を停止させるだけでなく，ハイエンドディスクドライブとモバイル向けディスクドライブを組み合わせることにより，より多様なシステム構成を可能とする．

一方で，近年のディスクドライブが有する多様な省電力化機能の活用を目指す研究も行われている．AutoMAID[8]は，ディスクドライブが有する低速回転アイドルモードを活用する商用ディスクアレイである．一般に，ディスクドライブはスピンドルモータを完全に停止させると，再びスピンアップするために，多くの電力量と時間を必要とするが，アイドル時に完全にスピンドルモータを停止させるのではなく，低速で回転させておくことにより，消費電力を低減するとともに，制御損の抑制を目指すものである．

なお，現状入手可能な商用ディスクドライブの一部では，アイドルモード時の回転速度の変更が可能であるが，磁気工学の進展により，回転速度を動的に変更可能なディスクドライブが近い将来に登場すると見られている．これらの先進的

デバイスの活用を目指す研究として，DRPM (Dynamic RPM)[5]では，ディスクストレージの回転速度をアクセス要求に応じて動的に制御する手法が提案されている．また，Hibernator[10]ではMAID及びPDCのアプローチと，DRPMのアプローチを組み合わせ，複数の回転速度モードを有するディスクドライブを活用して，ストレージ空間を複数のティア(tier)に分割し，ティア間でのデータブロック移送制御，並びにディスクドライブの回転速度制御によって省電力化を目指す．

本論文では，同期ログ転送方式に基づくディザスタリカバリシステムを対象として，二次サイトにおけるログ適用が非同期的に行われる点に着目して，データボリュームをアイドル化し，消費電力を削減する方式を提案している．限定された環境ではあるが，従来手法とは異なり，ストレージシステムと上位のデータベースシステムを連携させることにより，高い省電力化効果を達成可能としている点に特徴がある．

6. まとめ

ディザスタリカバリシステムにおける二次サイトの運用コストを低減することを目的として，本論文では，ログ転送方式による遠隔コピーを利用したディザスタリカバリシステムを対象に，ログ転送によりサイト間でデータ一貫性が保持される点に着目した省電力化方式を提案した．また，TPC-Cベンチマークを利用した評価実験においては，100[s]程度のRTO制約の下で，二次サイトにおけるディスクストレージの85%程度の省電力化が達成されることを示した．提案手法は，ディスクストレージの省電力化手法としては，一見極めて限定された環境を要求しているように見受けられる．しかし，ストレージシステムの多くが内部に冗長構成を有しており，一層の高い可用性が求められている背景からも，ディザスタリカバリシステムが今後広く利用されることは確実である．即ち，今後出荷されるストレージシステムのうち一定の割合が二次サイトで利用されることを鑑みるに，本研究が果たす約割は極めて大きいと言える．本論文では，ディスクストレージの消費電力として，主要コンポーネントであるディスクドライブを取り扱い，ディスクストレージを構成する電源，ファン，制御器に関する消費電力を無視した．本論文では，ボリューム単位，即ち比較的大きな粒度で資源の電力調整を行うため，これらの補助的コンポーネントも同様に制御可能であり，故に，本研究の評価における近似の影響は限定的であると考えているが，より詳細なモデルの構築を行い，確認を行いたい．

[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] British Standards Institution. BS25999: Business Continuity Management, 2006.
[2] 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.
[3] D. Colarelli and D. Grunwald. Massive Arrays of Idle Disks for Storage Archive. In Proc. Int'l Conf. on Supercomputing, pp. 1-11, 2002.
[4] Eagle Rock Alliance. Contingency Planning Research, 1996.
[5] 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.
[6] M. Ji, A. Veitch, and J.Wikes. Seneca: remote mirroring done write. In Proc. USENIX Conf. on File and Storage Tech., pp. 253-268, 2003.
[7] National Climate Data Center, U.S. DOC. Climate of 2005 Atlamtic Hurricane Season. Online Report available at http://www.ncdc.noaa.gov/oa/climate/research/2005/hurricanes05.html, 2005.
[8] Nexsan Technologies. Disk Based Storage Solutions: The Next Generation Now. Presentation Material, 2005.
[9] U.S. SEC. General Rules and Regulations promulgated under the Securities Exchange Act of 1934, 2005.
[10] 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.
[11] 合田和生, 喜連川優. データベース再編成機構を有するストレージシステム. 情報処理学会論文誌 : データベース, Vol. 46, No. SIG8(TOD 26), pp. 130-147, 2005.
[12] 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. 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), L4-2, 2007.
[13] 日立製作所. SANRISE 連携で実現する HiRDB ディザスタリカバリ構成の性能検証結果. ホワイトペーパー, 2004.
[14] 日立製作所. ログのみ同期転送で通信コストを削減する高信頼ディザスタリカバリ技術. はいたっく, Vol. 8, pp. 15-16, 2005.
[15] 渡辺聡, 鈴木芳生, 水野和彦, 藤原真二. ディザスタリカバリシステムにおけるログ適用処理の IO 回数削減手法の提案と評価. 情報処理学会北海道支部情報処理北海道シンポジウム, pp. 15-20, 2005.

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).