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

合田　和生†　　喜連川　優†

† 東京大学 生産技術研究所 〒 153-8505 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　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おり，業務継続を目的とした具体的な対策が，企業レベルだけでなく，国家レベルの法制度によって義務付けられつつある．特に，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を構築することが必須要件である．しかし，サイトレベルの冗長性を課す当該システムに関しては，必然的に倍化する運用コストの低減が極めて重要な課題である．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，ログ流量解析と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．

キーワード　データベースシステム，ディザスタリカバリ，ストレージシステム，トランザクション処理，遠隔コピー，省電力化

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

Kazuo GODA† and Masaru KITSUREGAWA†

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

Abstract　It has been widely recognized that the business breakdown due to unpredictable disasters such as terrors and hurricanes provides a nation and a society with terrible damage. Tangible actions intended for business continuity are being enforced by nation-level legal systems as well as by enterprise-level internal disciplines. Recent highly digitalized business cannot be operated without disaster recovery systems which incorporate a geographically distant secondary site as a backup system. However, for such a disaster recovery system involving site-level redundancy, a new crucial issue arises regarding the reduction of doubled operational cost. 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 log flow analysis and benchmark are disclosed, showing that two goals, power reduction of secondary site and quality preservation of business continuity, can be achieved. The contribution opens the gate for significantly reducing operational cost of disaster recovery systems.

Key words　Database System, Disaster Recovery, Storage System, Transaction Processing, Remote Copy, Power Reduction

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 [8]，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 [21].

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，企業レベルだけでなく，国家レベルの法制度によって義務付けられつつある．例えば，米国

SEC(Securiries and Exchange Comission) では，金融機関を対象とした一般規則 17 条 a 項 (Rule 17a.) [31] において，企業に重大な災害においても 1 営業日以内(注1)に業務を回復すること，広域災害へ対応可能とすること，並びに人的資源確保並びにシステム試験を義務づけている．同様に，我が国でも政府によって事業継続ガイドライン [40] が制定されつつあるほか，国際的にも英国 BS25999 [2](旧 PAS56 [1]) 並びに米国 NFPA1600 [22] を土台として，2010 年までに ISO における標準化が予定されている．

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 [7, 9, 10, 16, 18, 19, 25, 32, 42] 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，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，ログ流量解析と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するとともに，4. においては，3. の議論を踏まえ，ログ流量解析モデル，並びに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

2.1 ディザスタリカバリシステム

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．

2.2 同期転送と非同期転送

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．

同期転送は，一次サイト内における入出力処理において，データの書き込みが記憶装置に対してなされる際，一次サイト内の記憶装置に書き込まれたデータは即座に二次サイトに転送され，一次サイトの記憶装置，並びに二次サイトの記憶装置双方からの受領確認を以って，当該データ書き込み処理が成功したものと見なす方式である．当該方式では，常に一次サイトと二次サイトの記憶装置を完全に一致させることが可能であるため，ディザスタリカバリシステムに応用した場合，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，一次サイトの入出力処理において，書き込み時間に二次サイトへのデータ転送時間が含まれるため，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が発生する．

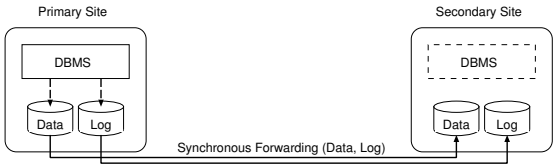
非同期転送は，同期転送とは異なり，一次サイト内における入出力処理において，データの書き込みが記憶装置に対してなされた際，一次サイト内の記憶装置からの受領確認のみを以って，当該データ書き込み処理が成功したものと見なす方式である．一般には，一次サイトの記憶装置への書き込みデータは，幾らかの遅延を以って，二次サイトへ転送され，二次サイトの記憶装置へ反映されることになる．当該方式では，一次サイトの入出力性能は，サイト間の距離に影響を受けない．一方，常に一次サイトと二次サイトの記憶装置上のデータは完全に一致していることを保証できないため，ディザスタリカバリシステムに応用した場合，災害時に二次サイトで業務を継続した際に失われる可能性を排除できない．

2.3 同期ログ転送

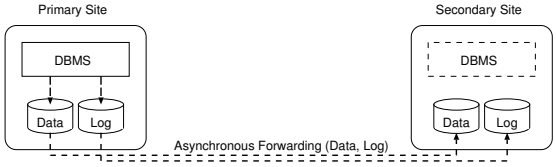
同期転送並びに非同期転送の双方の問題を解決するために，データベースシステムの記憶管理の特徴を利用した同期ログ転送 (Synchronous Log Forwarding: SLF) なる新しい方式が提案されている [43].

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

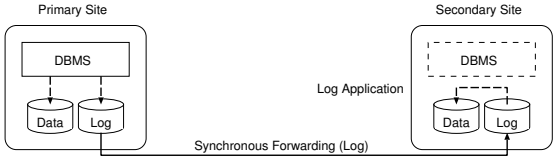
(注1)：将来は 2 時間から 4 時間以内を目指す．



(a) Synchronous Forwarding



(b) Asynchronous Forwarding



(c) Synchronous Log Forwarding

図 1　ディザスタリカバリにおける遠隔コピー技術

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．

図 1 にディザスタリカバリシステムにおける同期転送，非同期転送，並びに同期ログ転送をまとめる．

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

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(注2)．

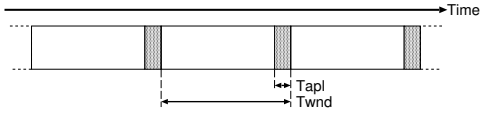


図 2　バッチ周期によるログ適用

3.2 同期ログ転送におけるログ流量解析

同期ログ転送によるディザスタリカバリシステムの基本的な変数として，以下の定数を定める．

R_{gen}: 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}: 二次サイトにおいて適用可能な単位時間あたりの最大ログレコード数
T_{RTO}: システムの許容する RTO

この際，同期転送されたログに関しては，二次サイトにおいて，図 2 に示すように一定の周期を以って非同期的に適用されるとする．当該ログ適用を規定するパラメータを以下にの通り定める．

T_{wnd}: 二次サイトにおけるログ適用のバッチ周期
T_{apl}: 二次サイトにおけるログ適用のバッチ周期においてログを適用する期間

二次サイトにおける各バッチ周期における時刻 t ($0 \leqq t < T_{wnd}$) において，それぞれ一次サイトから転送されたログ量 V_{fwd}，並びに二次サイトで適用したログ量 V_{apl} は以下の通りとなる．

$$V_{fwd}(t) = t \cdot R_{gen}$$

$$V_{apl}(t) = \begin{cases} 0 & (t < T_{wnd} - T_{apl}) \\ (t - T_{wnd} + T_{apl}) \cdot R_{apl} & (\text{otherwise}) \end{cases}$$

この際，ディザスタリカバリシステムが RTO 条件を満たすためには，以下の式を満足する必要がある．

$$\max_t \left(\frac{V_{fwd}(t) - V_{apl}(t)}{R_{apl}} + T_{up}(t) \right) \leqq T_{RTO}$$

ここに，$T_{up}(t)$ は時刻 t においてディスクドライブをアクセス可能な状態に遷移させるのに必要な時間を意味する．

上記の制約の下で，ディスクストレージの消費電力を最も削減可能な T_{wnd} 並びに T_{apl} は以下の通りとなる．

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

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

(注2)：本論文では紙面の都合上，議論の詳細は別稿に譲り，その概要を記するに留める．

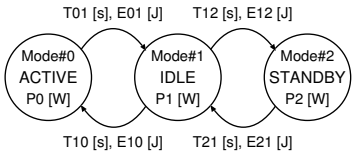


図 3　代表的な 3 モードディスクドライブ消費電力の状態遷移モデル

3.3　ディスクドライブの消費電力状態解析

近年のディスクドライブは，省電力化を目的として，複数の消費電力モードを有する．本論文では，図 3 に示す最も代表的な以下の 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以降の議論を行う．

モード 0 (アクティブ)　ディスクドライブにアクセス中の状態

モード 1 (アイドル)　ディスクドライブにアクセスはしていないが，入出力要求の到達によりモード 0 へ遷移することが可能な状態．

モード 2 (スタンバイ)　ディスクドライブの機械機構を停止し，消費電力を抑制している状態．

この際の記号を以下にまとめる．

P_i: 消費電力モード i における定常的な消費電力
E_{ij}, T_{ij}: 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

ただし，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．

前節の議論により，ディザスタリカバリシステムにおいては，二次サイトのデータボリュームに関して，ログ適用を周期的に実施するため，この際，バッチ周期 T_{wnd} 中のログ適用時にモード 0 へディスクドライブを遷移し，ログ適用終了時にモード 2 へディスクドライブを再度遷移させる．

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．

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

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．

3.4　有効性に関する考察

前節までに述べた簡易モデル化議論により，提案手法の有効性を定量的に把握することが可能となる．次章以降では，当該議論に基づき，ベンチマークを利用した定量的な評価結果を示すが，それに先立ち，本節では，以下の 2 点に関して，これまでに得られた観察から，有効性の考察を行う．

3.4.1　ログ生成レート R_{gen} とログ適用レート R_{apl} の比率

これまでの議論により，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 R_{gen} とログ適用レート R_{apl} の比率が寄与する割合が高い．即ち，$\frac{R_{apl}}{R_{gen}}$ 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．

$\frac{R_{apl}}{R_{gen}}$ を向上させるためには，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 [38] 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(注3)．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 IO 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．

図 4 に，TPC-C ベンチマーク [30] を用いた場合の，論文 [38] で述べたログ畳み込み，並びにログ整列によるログ適用の高速化の検証結果を示す．この場合，DBMS は 512MB のデータベースバッファを用い，TPC-C の思考時間は 0 とし，最大トランザクション処理スループットによってログが生成されている．対して，ログウィンドバッファとして同じ 512MB を用いた場合，ウェアハウス数 16，及び 160 の双方の場合において，約 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る．本実験では，最大トランザクション処理スループット時の R_{gen} を以って，高速化率を計測したが，実際のデータベースシステムでは，システムの最大スループットで定常的に運用することはないことから，$\frac{R_{apl}}{R_{gen}}$ はより高くなることが期待される．

二次サイトにおけるデータボリュームの省電力化のためには，$\frac{R_{apl}}{R_{gen}}$ を向上させる必要があるが，既存のログ適用高速化手法によっても，少なくとも 2 桁の高速化が可能であることが明らかになっている．なお，当該分野の研究としては，他にも論文 [39] などがあり，同様に著しい高速化が期待される．

3.4.2　データボリュームの省電力化によるディスクストレージ全体への寄与

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．

(注3)：著者らは，当該論文において，オンライン再編成を実現する目的を以って，データベース再編成の後におけるログ追い付きの高速化を議論した．データベース再編成後のログ追い付きは，ログレコードの参照アドレスを変換する必要があるなど，一般的なログ適用と比較してより複雑な処理となる．本論文で議論する同期ログ転送に基づくディザスタリカバリシステムのログ適用に応用できることは言うまでもない．

表 1　TPC-C ベンチマークにおける代表的なシステム構成

Rank	Vendor	System	tpmC	Database	# of disks (data volume)	# of disks (log volume)
1	IBM	System p5 595	4,033,378	IBM DB2 9	6400	360
2	IBM	eServer p5 595	3,210,540	IBM DB2 IDB 8.2	6400	140
3	IBM	eServer p5 595	1,601,784	Oracle Database 10g	3200	96
4	Fujitsu	PRIMEQUEST 540 16p/32c	1,238,579	Oracle Database 10g	1920	224
5	HP	Integrity Superdome Itanium2-64p/64c	1,231,433	Microsoft SQL Server 2005	1680	56

Quoted from *Top Ten TPC-C by Performance Version 5 Results* disclosed at `http://www.tpc.org/`, as of December 11, 2006.

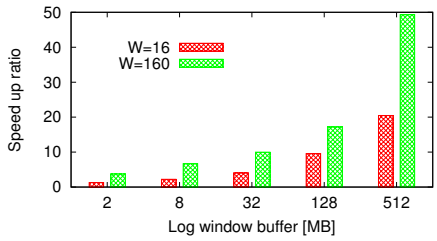


図 4　ログ畳み込み，及びログ整列によるログ適用の高速化の一例

表 1 に，TPC-C ベンチマークにおいて公表された性能上位のシステムに関して，その構成をまとめる．1 件を除き(注4)，その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，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をログ流量解析，並びにベンチマークを用いて評価する．

4.1 商用ディスクドライブの消費電力モデル

本論文では，既に論文 [11, 12] にて議論された IBM Ultrastar 36ZX，及び Fujitsu MHF 20043AT なる 2 つの商用ディスクドライブを例に，提案手法の有効性検証を行う．前者は主にサーバや大型ディスクアレイでの利用を想定したハイエンド機種であるのに対し，後者はラップトップ PC やモバイル環境での利用を想定したローエンド機種である．両ディスクドライブとも図 3 に示す 3 つの消費電力モードを有しており，表 2 に 3 つの消費電力モードの定常的な消費電力，並びに消費電力モード間の遷移遅延及び遷移エネルギーをまとめる．

4.2 ログ流量解析に基づく省電力化効果の見積り

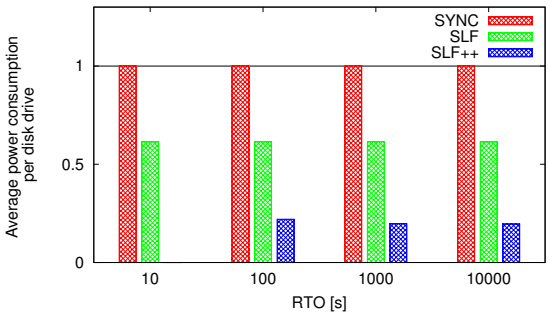
3. で行った議論に基づき，ディザスタリカバリシステムの二次サイトにおけるデータボリュームの省電力化効果を見積る．ここでは，以下の 3 つの場合を比較する．

SYNC　同期転送によるディザスタリカバリシステム．二次サイトのデータボリュームは常にアクセスされている状態である．

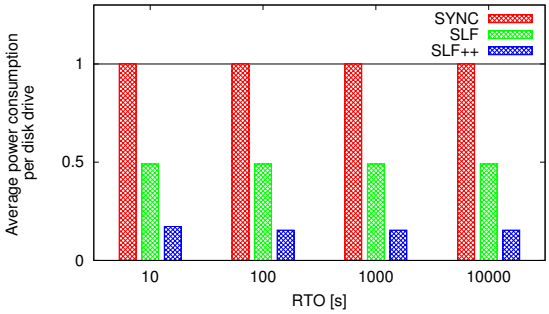
(注4)：順位 4 番目の Fujitsu 社製システムにおいても，データボリュームの占める資源は約 90%にのぼる．

表 2　商用ディスクドライブの消費電力モデルのパラメータ

Model	IBM Ultrastar 36ZX	Fujitsu MHF 20043AT
Steady-state power consumption		
P_0	39 [W]	2.2 [W]
P_1	22.3 [W]	0.95 [W]
P_2	4.15 [W]	0.13 [W]
Transition delay and energy		
T_{01}, E_{01}	0 [s], 0 [J]	0 [s], 0 [J]
T_{10}, E_{10}	0 [s], 0 [J]	0 [s], 0 [J]
T_{12}, E_{12}	15 [s], 62.25 [J]	0.67 [s], 0.36 [J]
T_{21}, E_{21}	26 [s], 904.8 [J]	1.6 [s], 4.4 [J]



(a) Power reduction effect on IBM Ultrastar 36ZX



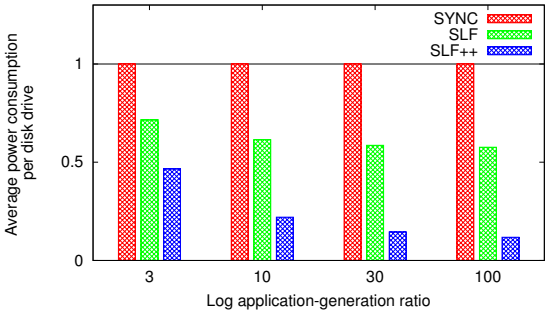
(b) Power reduction effect on Fujitsu MHF 20043AT

図 5　データボリュームの省電力化効果 ($\frac{R_{apl}}{R_{gen}} = 10$ の場合)

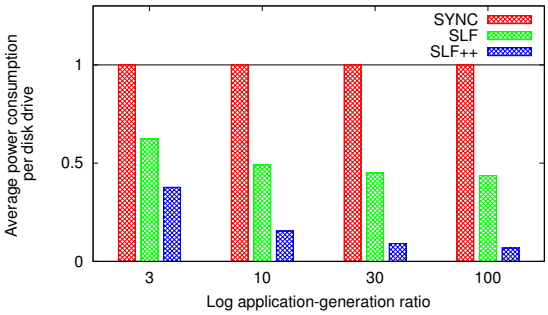
SLF　同期ログ転送によるディザスタリカバリシステム．二次サイトのデータボリュームは非同期的に実施されるログ適用時にのみアクセスされる．

SLF++　**SLF** の場合に加えて，3. で示したデータボリュームの省電力化制御を行うディザスタリカバリシステム．

先ず，図 5 に，T_{RTO} の変化による消費電力の変化をグラフとして示す．なお，ここでは比較のため，データボリュームの消費電力を，**SYNC** の場合をベースラインとした正規化により



(a) Power reduction effect on IBM Ultrastar 36ZX



(b) Power reduction effect on Fujitsu MHF 20043AT

図 6　データボリュームの省電力化効果 ($T_{RTO} = 100$ [s] の場合)

示す．IBM Ultrastar 36ZX における $T_{RTO} = 10$ の場合，**SLF++** の値を示していないが，これは，RTO 制約と比較して消費電力モード制御に係る時間が大きいため、データボリュームの消費電力モード制御が不可能であることを意味する，一方，それ以外の場合は全て，消費電力モード制御が可能である．

2 つの商用ディスクドライブの双方において，**SLF** によっても約 50%程度の省電力化が可能であることが分かる．これに対し，**SLF++**では，約 80%から 90%の大幅な省電力化を達成可能であることが分かる．RTO 制約が大きい程，若干省電力化効果は大きいものの，特段に RTO を大きくする必要性はなく，むしろ，**SLF++**によってディスクドライブの消費電力モード制御が可能である範囲，即ち，IBM Ultrastar 36ZX では凡そ 100 秒程度，Fujitsu MHF 20043AT では凡そ 10 秒程度の RTO が許容される場合，十分な省電力化が期待されることが明らかになった．

次に，図 6 に，ログ適用レートの高速化率 $\frac{R_{apl}}{R_{gen}}$ の変化による消費電力の変化をグラフとして示す．

SLF 並びに **SLF++**ともに，ログ適用を高速化することにより，省電力効果が高まっているが，その度合は **SLF** では凡そ 40%から 55%程度と限定的であるの対し，**SLF++**では大幅な省電力化を達成し，最大で 90%から 95%の省電力化を達成している．即ち，非同期的なログ適用における高速化技法と組み合わせることにより，提案手法が極めて有効に機能することが明らかになった．

4.3 TPC-C ベンチマーク環境を想定した省電力化効果の検証

TPC-C ベンチマーク [30] を用いて，ログ適用の高速化技法を利用した **SLF++**による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

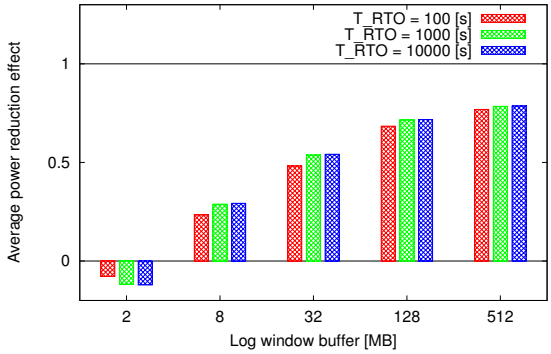


図 7　ディスクストレージの省電力化効果 (TPC-C ベンチマーク)

は，表 1 の順位 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前節の IBM Ultrastar 36ZX を利用した．ログ適用の高速化としては，著者らが論文 [38] で示した手法を商用 DBMS である HiRDB [41] 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．

図 7 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も，提案手法が有効に機能することが明らかになった．

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) [5,6] 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．キャッシュミス時に，コールド領域のドライブをアクセス可能な状態に移行させる (スピンアップする) 必要があるため，ミスペナルティが大きいが，当該方式は主に，三次記憶へのディスクストレージの応用として，Copan Systems によって商用化されるに至っている [20].

PDC (Popular Dara Concentration) [4, 27] 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 [29] は，ディスクドライブが有する低速回転アイドルモード [15] 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 [23, 33] 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) [12, 13] 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 [36] 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．

一方，従来の RAID 制御器を置き換える目的の提案，並びにキャッシュ管理に着目した提案が行われている．EERAID (Energy Efficient RAID) [17] は，ディスクストレージの RAID 制御器において，そのデータ冗長性に着目し，キャッシュ管理と入出力スケジューリングのアルゴリズムを改変することにより，ディスクストレージの省電力化を目指している他，RIMAC [34] では，同様に階層化された RAID 環境での省電力化を目指している．また，論文 [24] では，ディスクドライブのアイドル時間を長期化させることにより消費電力を最小化させる入出力のバースト化の提案がなされている．同様に，論文 [37] では，キャッシュ管理アルゴリズムによる省電力化効果の検証が行われている．

ストレージシステムの省電力化に関する研究が活発になりつつあることを受けて，検証のためのモデル化議論，並びに省電力化方式のシミュレーション技術に関する研究も行われている．Dempsey [35] では，ディスクドライブの消費電力を精密モデル化する試みがなされており，その成果は広く利用されているディスクドライブシミュレータである DiskSim [3] の拡張機能として実装されている．また，省電力化制御の有効性を検証するためのトレースドリブンな環境として，Dempsey を用いた Drive-Thru [26] が提案されている．

上記のストレージシステムの省電力化に関する研究は，主にストレージシステム内での制御方式に焦点が絞られている．これに対して，ディスクアクセスの主体であるサーバ上のアプリケーションやミドルウェアとの連携が模索されつつある．論文 [28] では，コンパイラの技法を用いた省電力化として，アプリケーションコードからディスクアクセスのプロファイルを得て，ストレージ上のデータのレイアウト変更を行う提案がなされている．また，論文 [11] では同様にアプリケーションコードから入出力パターンを抽出するが，この際，プログラムカウンタに着目したプロファイル手法を提案している．論文 [14] では，コンパイラにおいて入出力のバースト化を行うアプリケーション変換技法が提案されている．

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，その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，ログ流量解析とベンチマークを用いた評価実験を行い，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，より詳細なモデルを構築を行い，確認を行いたい．また，本論文では定常状態の流量モデルを用いたが，実際のトランザクション処理システムでは，過渡的な性能擾乱が発生する可能性があり，性能擾乱に対するシステム制御の安定性に関しても，今後，シミュレーションや実装を通じて評価を行う予定である．

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. PAS56: Guide to Business Continuity Managemenet, 2003.

[2] British Standards Institution. BS25999: Business Continuity Managemenet, 2006.

[3] 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.

[4] 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.

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

[6] 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.

[7] IBM Corp. DFSMS/MVS Version 1 Remote Copy Administrator's Guide and Reference, 1997.

[8] Eagle Rock Alliance. Contingency Planning Research, 1996.

[9] EMC Corp. Symmetrix Remote Data Facility product description guide, 2000.

[10] EMC Corp. EMC Enterprise SRDF consistency group: description and usage validation. Engineering White Paper, 2002.

[11] C. Gniady, Y. C. Hu, and Y-H. Lu. Program Counter-Based Prediction Techniques for Dynamic Power Management. *IEEE Trans. Comput.*, Vol. 55, No. 6, pp. 641–658, 2006.

[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*, Vol. 36, No. 12, pp. 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.*, pp. 121–130, 2002.

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

[16] 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.

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

[18] Dave McAuley. Planning for IBM Remote Copy. *IBM Redbooks*, 1995.

[19] Claus Mikkelsen and Tom Attanse. Addressing Federal Government Disaster Recovery Requirements with Hitachi Freedom Storage. White Paper, Hitachi Ltd., 2002.

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

[21] 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.

[22] National Fire Protection Association. NFPA1600: Standard on Disaster/Emergency Management and Business Continuity Programs (2004 Edition), 2004.

[23] 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*, pp. 2213–2215, 2000.

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

[25] H. Patterson, S. Manley, M. Federwisch, D. Hitz, S. Kleiman, and S. Owara. SnapMirror: File-System-Based Asynchronous Mirroring for Disaster Recovery. In *Proc. USENIX Conf. on File and Storage Tech.*, pp. 117–130, 2002.

[26] D. Peek and J. Flinn. Drive-Thru: Fast, Accurate Evaluation of Storage Power Management. In *Proc. USENIX Tech. Conf.*, pp. 251–264, 2004.

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

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

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

[30] Transaction Processing Performance Council. TPC-C, an online transaction processing benchmark. `http://www.tpc.org/tpcc/`.

[31] U.S. SEC. General Rules and Regulations promulgated under the Securities Exchange Act of 1934, 2005.

[32] Veritas Corp. VERITAS Volume Replicator 3.5: Administrator's Guide, 2002.

[33] 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*, Vol. 36, No. 5, pp. 2213–2215, 2000.

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

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

[36] 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.

[37] Q. Zhu and Y. Zhou. Power Aware Storage Cache Management. *IEEE Trans. Comput.*, Vol. 54, No. 5, pp. 587–602, 2005.

[38] 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. SIG 8(TOD 26), pp. 130–147, 2005.

[39] 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.

[40] 内閣府. 事業継続ガイドライン第一版–わが国企業の減災と災害対応の向上のために–, 2005.

[41] 日立製作所. Hitachi HiRDB Version 7. `http://www.hitachi.co.jp/Prod/comp/soft1/hirdb/`.

[42] 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.

[43] 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.