グリーンレプリケーション：二次系ディスクストレージの省電力化

合田　和生[†a)]　　喜連川　優[†b)]

Green Replication: Power Saving of Disk Storage for Secondary Systems

Kazuo GODA[†a)] and Masaru KITSUREGAWA[†b)]

あらまし　本論文では，グリーンレプリケーションと称し，業務継続を目的としたレプリケーションシステムにおける二次系ディスクストレージの省電力化方式を提案する．提案手法は，サービス復旧にかかる時間を意識した制御系のもとで，二次系に転送された更新情報をコンパクト化し，更新の反映操作を集中化することによって，ディスクドライブを長時間アイドル化する．商用データベースシステムを用いた実験により，30 秒から 100 秒程度のサービス停止時間のオーバヘッドのもとで，二次系ディスクストレージの消費電力のうち 80〜85%を削減可能であることを示す．

キーワード　グリーンストレージ，データベースレプリケーション，ストレージレプリケーション，データベースログレプリケーション，ディザスタリカバリ

1. まえがき

IT システムの電力消費は高い関心を集めている．増加する電力消費はシステムへの投資を圧迫し [1]，また，データセンタにおいては電力と熱の管理が，システム設計と運用を厳しく律速しつつある [2]．電力消費の削減策は IT システムに関する研究において重要な課題であり，この際にストレージシステムは無視できないサブシステムである [3]．SNIA は，典型的データセンタにおける全電力消費のうち，約 13%がストレージシステムによって直接消費され，また，配電設備や空冷設備への間接効果を含むとこれは約 27%にのぼると報告している [4]．ことに，データセンタの抱えるデータ量が爆発的に増大しており [5]，単にデータを貯蔵するだけでなくその処理スループットを向上するために，多数のディスクドライブがデータセンタに組み込まれるようになってきている．Q. Zhu ら [6] や M. Poess ら [7] の発表によると，大規模オンライントランザクション処理システムにおいては 7 割から 8 割程度の電力がディスクドライブによって消費されている場合があり，データインテンシブなエンタープライズシステムにおいて，ディスクストレージの省電力化は，サーバプロセッサやネットワーク装置の省電力化と同様に本質的な課題であるといえよう．

エンタープライズシステムにおけるディスクストレージの構成については，必ずしもすべてのディスクドライブがオンラインアプリケーションによって直接操作される一次データセットに使用されているわけではなく，むしろ，システムの可用性や性能を向上するために，より多くのディスクドライブが当該データセットの様々な複製を格納するために使用されている．例えば，同一のデータセンタ内に単一の物理スナップショットと，遠隔のデータセンタ内に単一のバックアップコピーを有する単純な IT システムにおいては，システム全体のディスクドライブのおよそ 3 分の 2 が通常，単に複製を格納するために利用されている．実際のエンタープライズシステムでは更に可用性を高めるためにバックアップを数世代用意するとともに，スナップショットを利用してバックエンドでデータベース解析や新たなソフトウェア開発を行うために，より多くのディスクドライブが主にこのような複製の格納に使用されている [8]．複製の格納を主に行うストレージ資源の省電力化は，一次データセットに使用されるストレージ資源の省電力化と比較して，オンラインの業務サービスとの分離性が高いため，実際のシステム

† 東京大学生産技術研究所，東京都
Institute of Industrial Science, The University of Tokyo, 4–6–1 Komaba, Meguro-ku, Tokyo, 153–8505 Japan
a) E-mail: kgoda@tkl.iis.u-tokyo.ac.jp
b) E-mail: kitsure@tkl.iis.u-tokyo.ac.jp

への導入の難易度が低く，省電力化ソリューションとして高い効果を有する可能性がある．

本論文では，系間での複製管理を行うレプリケーションシステムにおけるディスクストレージの省電力化手法を提案する．業務サービスの停止が，組織や場合によっては社会に与える影響は甚大であり [9], [10]，業務継続は組織レベルの内部統制のみならず国家レベルでの法規制によって義務づけられる場合もある [11], [12]．レプリケーションシステムは業務継続を支える重要なデータ保護基盤であり，例えば，ハードウェア故障やソフトウェアバグなどの障害に対してはデータセンタ内に二次系を用意し，また，テロやハリケーンといった災害に対しては遠隔地のデータセンタ内に二次系を用意し，一次系で業務サービスが停止した場合に迅速に二次系で当該サービスを再開することを可能とする．筆者らの提案では，レプリケーションのために転送される更新情報を二次系においてコンパクト化し，二次系のディスクストレージへのアクセスを集中化することにより，当該ディスクストレージの電力消費を大幅に削減することを目指す．省電力化によって，業務サービスの再開までにオーバヘッドが生じるものの，その影響は多くのシステムで許容できる範囲のものであると期待している．特に，遠隔地に二次系を設置するディザスタリカバリシステム [13]〜[15] の場合，現実には，一次系が正常に稼動している際に，二次系の多くの資源はほとんど利用されておらず，大幅な電力削減が期待される．筆者らは論文 [16], [17] においてディザスタリカバリシステムに関する検討を行ってきたが，本論文ではこれを一般のレプリケーションシステムに拡張して議論を行う．

本論文の構成は以下のとおりである．**2.** では本提案手法の想定として，多くのエンタープライズシステムで用いられているレプリケーションシステムを述べ，**3.** では当該レプリケーションシステムにおける電力消費の削減手法を明らかにする．**4.** では商用データベースシステムを用いた実験により提案手法の有効性を評価する．**5.** では関連研究を述べ，最後に **6.** で論文をまとめる．

2. レプリケーションシステム

本論文では，レプリケーションシステムにおける二次系ディスクストレージの省電力化を考えたい．高可用システムやディザスタリカバリシステムなどでは，通常時は業務サービスが一次系において稼動し，障害や災害が発生すると当該業務サービスは二次系において継続される．この際，業務サービスの復旧を可能とするためには，常に最新のデータを一次系から二次系に複製しておくことが不可欠であり，レプリケーションシステムはその基盤をなすものである(注1)．レプリケーションの基本的な方式は，一次系におけるディスクストレージ上のデータの更新情報を二次系に転送し，二次系において転送された更新情報をディスクストレージに反映するという二つのステップからなり，これまで多様なソリューションが論文などで提案され，また，実際のシステムに導入されてきた．今日のシステム構成では様々なレイヤでレプリケーションを行うことが可能である．例えば，データベースレプリケーションでは，一次系のデータベース管理システムが問合せ若しくはトランザクションを更新情報として二次系に転送し，二次系のデータベース管理システムが転送された問合せ若しくはトランザクションの処理を行う．また，最近のディスクストレージが有するサードパーティコピー機能においては，一次系から更新ブロックが更新情報として二次系に転送され，二次系においてディスクに反映される．更に，一部のデータベースベンダによってデータベースの更新ログを転送するソリューションが実用化されており，二次系において転送された更新ログをデータベースに適用することにより，レプリケーションを実現する．

このようなレプリケーションを基盤としたエンタープライズシステムの業務継続能力については，一般に，RPO (Recovery Point Objective) と RTO (Recovery Time Objective) なる二つの指標を以って議論することが多い．RPO はデータ損失の可能性であり，すなわち，一次系が停止した場合にシステムがどれだけ最近の更新データを二次系において復旧することを保証できるかを表す．一方，RTO は系間の引き継ぎにかかるオーバヘッドであり，一次系で業務サービスが停止した際にシステムがどれほど早期に二次系において当該サービスを再開できるかを表す．当然のことながら，RPO 及び RTO は双方とも小さいことが望ましい．特に，金融商取引などのエンタープライズシステムにおいてはわずかな時間で業務サービスを復旧することが求められ，かつ，いかなる障害若しくは災害

(注1)：このような単方向レプリケーションシステムのほかに，両系で独立に更新が行われるレプリケーションシステムも存在し論文等で議論が行われているが，今日のエンタープライズシステムにおいてはほとんど見られず，議論を別稿に譲りたい．

においてもデータを失うことは許されない．本論文では，以降の議論を簡単にするために，更新情報は同期的に転送され，よって RPO は常にゼロであるとする．

3. 省電力型レプリケーションシステム

本論文では，レプリケーションシステムにおいて，正常稼動時には複製の格納に資源のほとんどが利用されている二次系ディスクストレージの電力消費の削減方式を提案する．

図 1 に，議論の対象とするレプリケーションシステムを図示する．一次系において業務サービスが稼動し，業務サービスがディスクストレージに格納されたデータを更新する際に，当該更新情報が二次系に転送され，二次系においては当該更新情報がディスクストレージに反映される．この際，二次系のディスクストレージのうち，業務サービスのデータが格納される空間をメインボリュームと称し，一時的に更新情報が格納される空間をジャーナルボリュームと称することとする．ジャーナルボリュームは，例えばデータベースレプリケーションでは問合せバッファに，ストレージレプリケーションにおいてはいわゆるサイドファイルに，また，データベースログレプリケーションにおいてはログボリュームに相当するものである．なお，一次系で業務サービスが行われている際には，二次系の業務サービスは待機状態にあり，並行して他の処理が実行されることはないものとする．

3.1 更新情報の遅延化反映

筆者らの基本アイデアは，レプリケーションシステムにおける更新情報の反映が有する非同期性に着目し，更新情報のディスクストレージへの反映を遅延化することにより，二次系におけるメインボリュームへのアクセスを集中化し，当該ボリュームをスタンバイ化して消費電力を削減する機会を生み出すことにある(注2)．図 2 に，このような遅延化反映を実現するバッチ反映手法を図示する．二次系は，転送された更新情報を一時的にジャーナルボリュームへ格納し，この際，即座にメインボリュームに更新情報を反映するのではなく，バッチ間隔 (T_{wnd}) を設け，一括して反映を行う．すなわち，バッチ間隔のうち，最初の T_{def} の時間はメインボリュームに反映せず，その後の T_{apl} の時間に集中的にメインボリュームに反映する．反映が行われていない間，メインボリュームはアイドルであることから，T_{def} がメインボリュームのブレークイーブン時間(注3)より長い場合，当該ボリュームをスタンバイ化し，すなわち，ボリュームを構成するディスクドライブをスピンダウンすることにより，消費電力を削減することが可能である．遅延化時間 T_{def} をより長くとることにより，電力の削減効果は向上するが，同時に，ジャーナルボリュームにはより多くの未反映の更新情報が蓄積される．障害や災害のために二次系で業務サービスを開始する際には，一般に，二次系はまずすべての未反映の更新情報を反映する必要がある(注4)．よって，レプリケーションシステムの二次系においては，更新情報の反映にかかる最大の遅延化時間 T_{def}^{MAX} は，省電力化のために許容し得る RTO オーバヘッド T_{RTO} によって制約される．

RTO オーバヘッドと遅延化時間の関係に関して議論する．ここでは，一次系においては単位時間当り R_{gen} の更新情報が生成されるものとし，二次系においては単位時間当り R_{apl} の更新情報を反映することができるものとする．なお，レプリケーションシステ

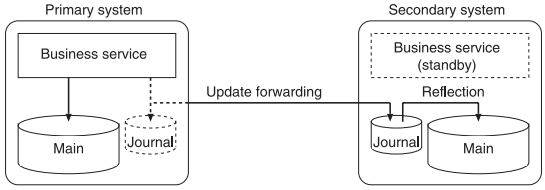


図 1 業務継続を目的としたレプリケーションシステム
Fig. 1 A replication system for business continuity.

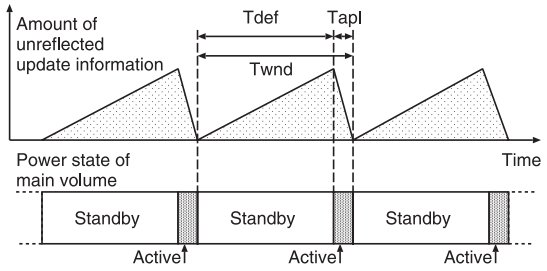


図 2 更新情報の遅延化反映
Fig. 2 Deferred reflection of update information.

(注2)：同様の効果は，例えば，一般的なデータベースシステムのバッファ管理における遅延化書込みにおいて，チェックポイント間隔を長く設定することによっても得られる場合があるが，トランザクション到着率が十分に低く，バッファ空間が十分に大きい場合に限定される．
(注3)：スタンバイ化による省電力化効果が，スタンバイ化とアクティブ化のオーバヘッドと拮抗するアイドル時間をブレークイーブン時間と呼ぶ [18]．
(注4)：すべての更新情報を反映する前に部分的に業務サービスを開始する技法 [19] も存在するが，本論文では議論の対象としない．

ムが機能するためには，$R_{apl} > R_{gen}$ であることが必要である．1 回のバッチ間隔において，二次系に到達する更新情報量は，二次系が反映可能な更新情報量以下である必要があることから，更新情報の遅延化時間 T_{def} とその後の集中的な反映時間 T_{apl} は，以下の関係を満たす必要がある．

$$(T_{def} + T_{apl}) \cdot R_{gen} \leq T_{apl} \cdot R_{apl}$$

また，任意の時点において，T_{RTO} 以内に，必要に応じてメインボリュームをアクティブ化し，すべての未適用の更新情報をメインボリュームに反映可能であることを保証する必要がある．遅延化時間がブレークイーブン時間より長いとし，遅延化の開始と同時にメインボリュームのスタンバイ化を開始し，また，遅延化の終了と同時にちょうどアクティブ化が完了するように制御できるものとすると，系は以下の関係を満たす必要がある．ただし，以下で，時刻 t はバッチ間隔が開始した起点を 0 とするものであり，T_{down}，T_{up} はそれぞれメインボリュームのスタンバイ化，アクティブ化にかかる時間を表す．

（時刻 0 にスタンバイ化を開始してから，時刻 T_{down} にスタンバイ化を完了するまでの任意の時刻 t について）

$$\frac{t \cdot R_{gen}}{R_{apl}} + (t - T_{down}) + T_{up} \leq T_{RTO}$$

（時刻 T_{down} にスタンバイ化を完了してから，時刻 $T_{def} - T_{up}$ にアクティブ化を開始するまでの任意の時刻 t について）

$$\frac{t \cdot R_{gen}}{R_{apl}} + T_{up} \leq T_{RTO}$$

（時刻 $T_{def} - T_{up}$ にアクティブ化を開始してから，時刻 T_{def} に反映を開始するまでの任意の時刻 t について）

$$\frac{t \cdot R_{gen}}{R_{apl}} + (T_{def} - t) \leq T_{RTO}$$

（時刻 T_{def} に反映を開始し，時刻 $T_{def} + T_{apl}$ に反映が終了するまでの任意の時刻 t について）

$$\frac{t \cdot R_{gen} - (t - T_{def}) \cdot R_{apl}}{R_{apl}} \leq T_{RTO}$$

上式により，系において遅延化時間 T_{def} の最大値 T_{def}^{MAX} と，その際に必要な反映時間 T_{apl}^{MAX} 並びにバッチ間隔 T_{wnd}^{MAX} は，以下のとおりとなる．

$$T_{def}^{MAX} = \frac{R_{apl}}{R_{gen}} \cdot T_{RTO}$$

$$T_{apl}^{MAX} = \frac{R_{apl}}{R_{apl} - R_{gen}} \cdot T_{RTO}$$

$$T_{wnd}^{MAX} = T_{def}^{MAX} + T_{apl}^{MAX} = \frac{R_{apl}^2}{(R_{apl} - R_{gen}) \cdot R_{gen}} \cdot T_{RTO}$$

ただし，$T_{RTO} > T_{down} + T_{up}$ でなければならない．

上記の遅延化パラメータに基づき更新情報の反映を集中化し，同時に能動的にデータボリュームのアクティブ化とスタンバイ化を制御することが可能となる．すなわち，二次系における制御系は，一次系から転送される更新情報のレートに基づき生成レート R_{gen} を，また，二次系において実現可能な更新情報の反映スループット R_{apl} を観測し，これらに基づき，最大の遅延化時間 T_{def}^{MAX} を求める．T_{def}^{MAX} がメインボリュームのブレークイーブン時間より長い場合には，当該ボリュームをスピンダウンすることにより消費電力の削減が期待されるため，更新情報の反映の遅延化を開始すると同時に，スタンバイ化を指示する．また，T_{def}^{MAX} に基づき，更新情報の反映を開始するに先立って事前にメインボリュームのアクティブ化を行うことが可能である．このように能動的にメインボリュームの電力モードを制御することによって，効率的にディスクストレージの消費電力を削減することが可能となる．

3.2 更新情報のコンパクト化

前節によって，$\frac{R_{apl}}{R_{gen}}$ が大きいほど，T_{def}^{MAX} が大きくなり，省電力化効果が高くなるため，更新情報の反映スループットの向上が鍵となることが分かる．本節では更新情報の反映を高速化するためのコンパクト化技法を示す．

一般にレプリケーションシステムにおいて，二次系に転送された更新情報は，データベースの REDO オペレーションに類似する操作によって，メインボリュームに反映される．すなわち，ジャーナルボリュームにおいては，更新情報のエントリは当該エントリが生成された順序で記録されていることから，ジャーナルボリュームから未反映の更新情報のエントリを記録順序に従って読み出し，当該エントリを読み出した順序で逐次的にメインボリュームに反映することによって，更新情報をそれが記録された順序に従って反映する．これに対して，本論文で提案する手法では，ジャーナルボリューム中に記録された未反映の更新情報のうち，最古のエントリから順に一定量の更新情報を一括して

主記憶上に用意したコンパクト化用バッファに読み込み，当該更新情報をコンパクト化用バッファを用いてコンパクト化し，その後にコンパクト化された更新情報を一括してメインボリュームに反映する．コンパクト化操作は，更新情報の集約と更新情報の整列なる二つの手法から構成される．

更新情報の集約は，コンパクト化用バッファ中の更新情報において同じデータを更新するエントリが複数存在する場合に，当該エントリを畳み込むことにより，更新情報のエントリ数を削減する．例えば，データベースログレプリケーションの場合，`insert(data1)`，`update(data1 → data2)`，`update(data2 → data3)` なる同じレコードを更新する三つのエントリ系列は，単一のエントリ `insert(data3)` に集約することが可能であり，これにより反映コストを削減することが期待される [20]．同様に，ストレージレプリケーションにおいてはブロック単位で集約することが可能であろう．

一方，コンパクト化用バッファ中の更新情報において異なるデータを更新する各々のエントリに関しては，その適用順序を入れ換えることが可能であり，これにより，反映に掛かる入出力負荷を軽減することができる場合がある．更新情報の整列は，当該効果を得るために，コンパクト化用バッファ中で異なるデータを更新する更新情報のエントリの順序を入れ換えることにより，反映に掛かる入出力負荷を軽減するものである．データベースログレプリケーションやストレージレプリケーションの場合，更新情報中の各エントリは更新先の物理アドレスを情報として有しており，当該情報を活用してエントリの整列を行うことにより，入出力系列の大幅な改善が期待される．これらの技法を活用することにより，更新情報の反映スループットを向上し，よって電力削減効果の向上が期待される．

データベースログレプリケーションやストレージレプリケーションの場合，上記のとおり，更新情報が有する物理情報を生かしたコンパクト化操作が可能であるが，一方で，データベースレプリケーションの場合，更新情報である問合せやトランザクションは物理アドレス非依存な記述がなされるため(注5)，若干の変更が必要となる．筆者らは，バッチ問合せスケジューリング技法 [21], [22] を用いて問合せやトランザクションをスケジュールすることにより，近い効果が得られると考えている．

このような更新情報のコンパクト化技法は，むしろこれまでは，系間の通信トラヒックの軽減などの目的で検討され，主に非同期的なレプリケーションシステムにおける一次系で行われてきた [15] が，対して，本論文では，当該手法を二次系に適用し，ディスクストレージの省電力化を試みる点が新しいといえよう．なお，本論文では更新情報の集約並びに整列は待機状態にある二次系で実施されるものとしているため，並行して他の処理が実行されることはなく，システムの一貫性が崩れることはない．二次系において並行して他の処理を実行する場合の議論は，別稿に譲りたい．

3.3 更新情報とストレージ空間全体の関係

二次系のストレージシステムについては，当該ストレージ空間をメインボリュームとジャーナルボリュームに分けて議論してきたが，本節では，両ボリュームを含めた二次系のストレージシステム全体に対する提案手法の有効性を検討したい．

表 1 に，2009 年 10 月現在の TPC-C ベンチマークで公開された性能上位 5 件のシステムを示す．このうち，四つのシステムにおいては，すべてのディスクドライブのうち 2.1〜6.3%のディスクドライブをデータベースの更新ログを格納するログボリュームに，もう一つのシステムでも 10.0%をログボリュームに用いており，残りをデータボリュームに用いている．これは，オンライントランザクション処理システムなどの多くのデータインテンシブアプリケーションにおいては，おおむねランダムアクセスとなるデータボリュームのスループットを向上するために，非常に多数のディスクドライブを並列駆動する設計がなされており，その更新情報である更新ログは一般にシーケンシャルアクセスで書き込まれるため，比較的少数のディスクドライブで構成されたログボリュームにより十分にスループットを吸収可能であることを意味している．

レプリケーションシステムの二次系においても同様であり，少数のディスクドライブを以ってジャーナルボリュームを構成し，残りの多数のディスクドライブを以ってメインボリュームとするのが一般的である．ジャーナルボリュームを常にアクティブとしつつ，メインボリュームを相当時間スタンバイ化することを可能とする提案手法は，二次系のストレージシステム全体に対して高い省電力化効果を有すると期待される．

(注5)：更新ログが同様に物理アドレス非依存な形式で記録される場合におけるデータベースログレプリケーションについては，データベースレプリケーションと同様に議論することができる．

表 1　典型的なオンライントランザクション処理システム
Table 1　Typical online transaction processing systems.

順位	ベンダ	システム	tpmC	データベース	ディスクドライブ数（データボリューム）	ディスクドライブ数（ログボリューム）
1	IBM	Power 595 Server	6,085,166	IBM DB2 9.5	10752	240
2	HP	Integrity Superdome 64p/128c	4,092,799	Oracle Database 10g R2	6608	448
3	IBM	System p5 595	4,033,378	IBM DB2 9	6400	360
4	IBM	eServer p5 595	3,210,540	IBM DB2 UDB 8.2	6400	140
5	Fujitsu	PRIMEQUEST 580A 32p/64c	2,382,032	Oracle Database 10g R2	3456	384

2009 年 10 月 4 日時点．http://www.tpc.org/に公開された Top Ten TPC-C by Performance Version 5 Results より引用．
ディスクドライブ数はスペアを除く．

4. 評価実験

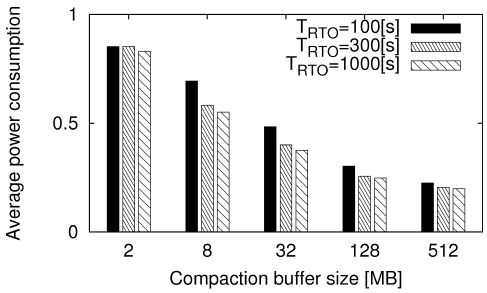
筆者らは提案手法の有効性を評価する実験を行った．実験においては，アプリケーションとしてオンライントランザクション処理を模した TPC-C ベンチマークを用い，データベースログレプリケーションによるレプリケーションシステムが構築されているものと想定した．オンライントランザクション処理はエンタープライズにおける代表的なデータインテンシブアプリケーションであり，そのデータは多くのエンタープライズにおいて業務上極めて重要であることから，しばしばデータセンタ内のローカルレプリケーションシステムやデータセンタ間のディザスタリカバリシステムによる保護の対象となっている．提案手法を当該レプリケーションシステムに適用することにより，業務継続の品質をほとんど低下させることなく，二次系におけるディスクストレージの消費電力を大幅に削減することが可能であることを示す．

実験に際しては，提案手法による潜在的な消費電力の削減を測定するために，ハイブリッドのシミュレーション環境を構築した．ディスクドライブモデルに基づき電力消費を算出することのできるソフトウェア型ディスクドライブシミュレータを用いるとともに，当該シミュレータの上で，コンパクト化技法を具備する更新情報の反映機構として，高度ログ適用器 [20] を実装した．高度ログ適用器は商用データベースシステムである HiRDB [14] によって生成されたデータベースの更新ログを適用することが可能である．

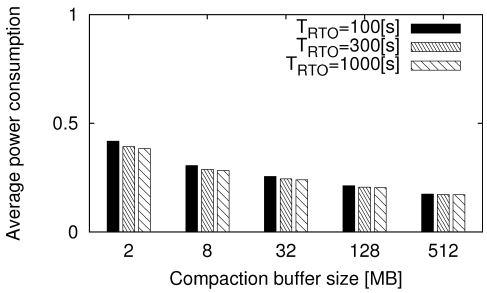
測定実験を二つの Xeon プロセッサと 2 GByte の主記憶を有する Linux サーバ上で行った．この際，ウェアハウス数を 16 並びに 160 とする TPC-C ベンチマークを構成し，それぞれの構成のもと，512 MByte のデータベースバッファを有する HiRDB 上で 100 万トランザクションを実行し，データベースの更新ログを生成した．この際同時に，カーネルレベルの入出力トレーサを用いて入出力挙動をトレースした．次に，シミュレーション環境においてディスクドライブモデルを用いてトレースした入出力を再生することにより，一次系における更新情報の生成，すなわち更新ログの生成をシミュレートした(注6)．その後，高度ログ適用器を用いて生成された更新ログを適用し，二次系における電力消費の削減効果を測定した．実際のレプリケーションシステムでは，一次系から二次系への更新情報の転送が通信路媒体等によって制約されることが多い．例えば，ディザスタリカバリシステムでは，系間の通信応答時間が一次系の処理性能をたびたび律速するが，本実験では，レプリケーションシステム一般に対する有用性を評価するために，二次系における更新情報の反映性能に注目した検証を行うこととし，一次系から二次系への更新情報の転送にかかる応答時間並びに帯域制約を無視した．なお，本実験を通じて，ディスクストレージの構成としては，表 1 における IBM System p5 595 のものに従った．よって，一次系及び二次系の双方において同一の構成を有するディスクストレージが用いられ，全ディスクドライブのうち，約 94%がデータボリュームに，残りがログボリュームに用いられるものと仮定した．実験においては，RTO オーバヘッド制約並びにコンパクト化用バッファサイズを変化させて測定を行った．なお，有効性を評価するために，提案システムである省電力型レプリケーションによる二次系ディスクストレージの消費電力を，二次系に転送された更新情報を一次系と同じ 512 MByte のデータベースバッファを用いて即座に反映する従来型レプリケーションシステムによる消費電力と比較した．

図 3 にハイエンドのディスクドライブモデルで得られた結果を示す．モデルの基本パラメータとしては

(注6)：100 万トランザクションは思考時間を 0 として発行し，よって一次系においては与えられたディスクドライブモデル上で，最大レートでトランザクションが処理されているものと仮定している．



(a) 16 warehouses



(b) 160 warehouses

図 3　ハイエンドディスクドライブを用いた二次系ディスクストレージにおける省電力化効果
Fig. 3　Power saving of secondary-site disk storage with high-end disk drives.

表 2　ディスクドライブモデルの基本パラメータ
Table 2　Basic parameters of disk drive models.

モデル	IBM Ultrastar 36Z15	HGST Deskstar T7K250
物理容量	18.4 GByte	250 GByte
回転スピード	15000 rpm	7200 rpm
平均シーク時間	3.4 ms	8.5 ms
転送レート	55 MByte/s	61 MByte/s
アクティブ時電力消費	39.0 W	9.7 W
アイドル時電力消費	22.3 W	5.24 W
スタンバイ時電力消費	4.15 W	(U) 4.04 W (L) 2.72 W (N) 0.93 W
スピンダウン時の時間損とエネルギー損	15.0 s 62.25 J	(U) 0.7 s, 3.5 J (L) 17.0 s, 19.0 J (N) 0.7 s, 3.5 J
スピンアップ時の時間損とエネルギー損	26.0 s 904.8 J	(U) 0.6 s, 3.0 J (L) 4.0 s, 37.9 J (N) 3.5 s, 107.0 J

(U): アンロードモード，(L): 低 RPM モード，
(N): ノンスピンモード

表 2 に示す IBM 製 Ultrastar 36Z15 のものを用いた．当該モデルは必ずしも最新のものではないものの，多くのこれまでの論文 [6], [23]〜[25] で用いられている信頼性の高いものであり，他に同様のモデルが見当たらないことから，まずこれを用いた．図では，各バーが二次系におけるディスクストレージの平均消費電力を示しており，これは従来型のシステムに対して正規化されたものである．

表 3　100 秒の RTO オーバヘッド制約下におけるハイエンドディスクドライブに対するバッチ間隔
Table 3　Batch intervals for high-end disks under 100 seconds of RTO overhead requirement.

ウィンドウバッファサイズ	バッチ間隔	
	16 ウェアハウス設定	160 ウェアハウス設定
32 MByte	536 s	1110 s
128 MByte	1070 s	1830 s
512 MByte	2150 s	5030 s

コンパクト化用バッファサイズを大きくすることにより，更新情報の反映スループットが向上し，これにより，より多くの電力を削減することができた．一次系におけるトランザクション処理には 512 MByte のデータベースバッファを用いたが，二次系において同サイズのコンパクト化用バッファを更新情報のコンパクト化に用いることにより，二次系のディスクストレージ全体で約 85%の消費電力を削減することができたことは注目に値する．一次系の更新情報の生成レートに対する二次系の更新情報の反映レートの比率 $\frac{R_{apl}}{R_{gen}}$ は，コンパクト化技法によってウェアハウス数 16 のもとでは 20.5 に，ウェアハウス数 160 のもとでは 49.3 に加速しており，この結果，例えば，100 秒という RTO オーバヘッド制約のもとでも，表 3 に示すとおり，バッチ間隔 T_{wnd}^{MAX} をそれぞれ 2150 秒，5030 秒とすることができ，二次系のディスクストレージのメインボリュームを相当時間，スピンダウンすることができるようになり，省電力化に至ったといえよう．一方，RTO オーバヘッド制約については，許容時間の拡大は省電力化効果の向上につながるものの，その効果は必ずしも大きくないことが分かった．本実験においては，30 秒という厳しい RTO オーバヘッド制約のもとでは，許容時間そのものがボリュームのアクティブ化及びスタンバイ化の時間損より短いことから省電力化の効果は得られなかったが，100 秒以上のより控え目な RTO オーバヘッド制約のもとでは大幅な消費電力の削減を実現することができた．

なお，上記の実験において，更新情報のコンパクト化を行わない場合に，どれほど更新情報の反映を遅延化できるかを検証したところ，$R_{apl} < R_{gen}$ となり，レプリケーションシステムとして系が機能しない結果となった．本実験では，システムの最大レートでトラ

ンザクションを注入しており，基本的にディスクストレージバウンドである．512 MByte のデータベースバッファを用いている一次系に対して，コンパクト化を行わない場合，二次系のストレージアクセス負荷が一次系のそれを上回るものとなった．特に，トランザクションの到着率が高い場合には，更新情報の遅延化反映のために更新情報のコンパクト化が欠かせないことが確認された．

次に，バッチ間隔 T_{wnd}^{MAX} について検討したい．ディスクストレージを構成するディスクドライブは機械的機構を有するため，頻繁な電力モードの変更はドライブの寿命に影響する可能性がある．表 3 に，先の実験におけるいくつかのケースの T_{wnd}^{MAX} 値を示す．小さいコンパクト化用バッファの場合，データボリュームの電力モードを頻繁に変更する必要があるものの，512 MByte のコンパクト化用バッファサイズの場合は，16 並びに 160 ウェアハウス設定でそれぞれ一日に約 40 回，17 回のモード変更（スピンアップとスピンダウンの組合せを 1 回とする）で収まっている．多くのハイエンドディスクドライブでは，おおむね 5 年程度の寿命が期待されることと，データシート上約 5 万回のモード変更が保証されていることを鑑みると，一日に約 40 回，17 回というモード変更の頻度から保証される製品寿命はそれぞれ約 3.4 年，8.1 年であり，前者の 16 ウェアハウス設定の場合，必ずしも十分なものとはいえない可能性がある．この点に関しては，今後検討を深めたいと考えているが，本実験では，一次系におけるトランザクション処理は常時最大レートで行われていると仮定されていることから，現実のシステムに適用した場合，それほど大きな問題とはならないものと期待している．当然のことながら，より大きな RTO オーバヘッドを許容した場合，バッチ間隔はより長くなる．更新情報のコンパクト化が省電力化効果だけではなく，ディスクドライブの製品寿命にも大きな影響を有することが明らかとなった．

更に，ニアラインストレージ用途に近年利用がなされており，様々な新しい省電力化機能 [26] を有するミッドレンジのディスクドライブを用いて同様の実験を行った．同じく，表 2 にモデルの基本パラメータを示す．当該ディスクドライブに関しては，十分なモデルパラメータが公開されているわけではなく，筆者らが論文 [27] で報告した手法を用いて計測したものである．HGST 製の T7K250 をベースとしたモデルは，アンロード，低 RPM，並びにノンスピン（従前の省電力化機能におけるスタンバイに相当）の三つのスタンバイモードを有する [26]．実験を行ったところ，提案手法は，基本的にはハイエンドのディスクドライブと同様に，このようなミッドレンジのディスクドライブにおいても有効に機能した．ただ，近年のミッドレンジのディスクドライブはハイエンドのディスクドライブと比べてより小さい時間損で電力モードを変更することができるため，30 秒という厳しい RTO オーバヘッド制約下においても省電力化を実現することが可能となった．図 4 に 160 ウェアハウス設定におけるこれらの三つのスタンバイモードを比較する．ノンスピンモードを用いることにより，従来型のレプリケーションシステムと比較して最大で約 80%の消費電力を削減することができた．一方で，本実験においては，アンロードや低 RPM といった新しい省電力下モードについて顕著な有効性を示すことはできなかった．

本章においては，TPC-C ベンチマークとレプリケーションシステムを想定した環境において，提案手法によって，業務回復能力をほとんど劣化させることなく，二次系ディスクストレージの消費電力を大幅に削減することができることを示した．ハイエンドのディスクドライブ及びミッドレンジのディスクドライブについては，それぞれ RTO について 100 秒及び 30 秒のオーバヘッドを認めることにより，おおむね 80〜85%の省電力化が実現できることが分かった．一次系における大規模災害などに備えてデータセンタ間でリモートレプリケーションを行うディザスタリカバリシステムについては，最も高い保護レベルとしてファイブナイン (99.999%) と称される可用性目標を掲げることがあり，これは年当り 315 秒のサービス停止が認められてい

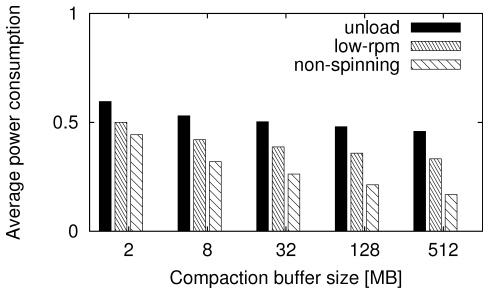


図 4　30 秒の RTO オーバヘッド制約下におけるミッドレンジディスクドライブを用いた二次系ディスクストレージにおける省電力化効果

Fig. 4　Power saving of secondary-site disk storage with mid-range disk drives under 30 seconds of RTO overhead requirement.

ることに相当する．30 秒若しくは 100 秒程度の RTO オーバヘッドを許容することは，比較的多くのシステムに対して受け入れやすいものであるだろう．一方，データセンタ内でローカルレプリケーションを行ういわゆる高可用システムについては，ハードウェア故障やソフトウェアバグに備えてより頻繁に一次系から二次系への業務サービスの引き継ぎが想定される場合があり，特に高い可用性を目指すシステムにおいては 10 秒以内の復旧が要請される場合がある．このような超高可用用途のローカルレプリケーションシステムについては，本提案技術によって十分な省電力化を期待することは難しいと考えられるものの，30 秒若しくは 100 秒程度の RTO オーバヘッドが受け入れられる場合においては，ディザスタリカバリシステムと同様に応用が期待されるといえよう．

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．最も単純なアプローチは，一定時間，ディスクドライブにアクセスがない場合に，当該ディスクドライブを低電力モードに移行させるものである．このような技法は広く商用のディスクドライブに採用されている．また，より洗練された技法として，低電力モードに移行するためのしきい値を適応的に制御する試み [28], [29] も行われてきている．このようなしきい値に基づく省電力化技法は，主に対話的な処理が支配的であって応答性能に対するユーザの許容が見られるエンドユーザ環境で有効なもので，バッテリ駆動型のモバイルコンピューティング環境では高い効果が見られる．一方で，高い応答性が問われるエンタープライズシステムにおいては，ドライブのスピンアップ時のオーバヘッドが許容されにくく，当該技法を直接適用することは難しいといえよう．

Massive Array of Idle Disks (MAID) [30] 並びに Popular Data Concentration (PDC) [24] は，頻繁にアクセスされるブロックを特定のディスクドライブに移送若しくは複製することにより，当該ディスクドライブにアクセスを集中させ，他のディスクドライブにおいてより長いアイドル時間を生成し，スピンダウンの機会を生み出すものである．このようなデータアクセスの局所性を活用する技法は，実際のアーカイブストレージシステムに採用されるに至っている．

近年のディスクストレージに搭載された RAID 機能が有する広大なキャッシュ空間とディスクドライブ上に生成される冗長情報を活用する手法も提案されている．Energy Efficient RAID (EERAID) [23] 及び RIMAC [25] は，スピンダウン中のディスクドライブに本来格納されているブロックのキャッシュからのエビクションをなるべく避けるように，RAID コントローラにおいて入出力要求を調整するものである．Power-Aware RAID (PARAID) [31] は，RAID-5 編成の上に非対称的なパリティレイアウトを導入し，アクティブなディスクドライブ数を動的に変更するものである．

回転速度を動的に変更可能な多段速のディスクドライブについて研究を進めているグループもある [6], [32]．このような試みは，実現されれば大変有効であろうが，筆者らの知る限り，そのような多段速のディスクドライブは実験室でのプロトタイプが報告されている [33], [34] に留まっており，商用化は予定されていない．

また，アプリケーション支援によるディスクストレージの省電力化アプローチも報告されている．Co-operative IO [35], [36] は，消費電力を意識した入出力システムコール群であり，アプリケーション開発者が個々の入出力に対してその遅延化可能性と中断可能性を指定し，これを利用してオペレーティングシステムが入出力発行を集中的に行う．アプリケーション変換手法 [37]〜[39] は，ソースコードのコンパイル過程において入出力命令を調整し，同じく入出力要求を集中化する．

近年では，deduplication [40] と称されるディスクストレージ上の重複ブロックを削減する技法が登場しており，当該技法を利用してディスクドライブ数を削減するソリューションが提案されている．当該手法はアーカイブ目的のストレージには非常に有効であろう．一方，本論文で議論するような業務継続を目的とするレプリケーションシステムでは，障害や災害が生じた際に二次系で業務サービスを提供することが期待され，一次系と二次系は同等の入出力帯域を確保することが必要であることから，直接の適用は困難であろう．

これらのアプローチに対して，本論文では，レプ

リケーションシステムにおける二次系のディスクストレージの特性を活用して，その省電力化を図るものである．当該ディスクストレージは，一次系で業務サービスが稼動している最中は，通常，複製の管理のみを行っており資源はほとんど利用されておらず，よって，更新情報の反映の非同期性を活用して，RTO を意識した制御系のもとで反映操作を遅延化させることにより，多くのディスクドライブを長時間アイドル化することが可能となり，著しい省電力化効果が期待される．

6. むすび

本論文では，グリーンレプリケーションと称し，業務継続を目的としたレプリケーションシステムにおける二次系ディスクストレージの電力消費の削減方式を議論した．コンパクト化技法を用いて二次系に転送された更新情報の反映を遅延化することにより，大幅な省電力化効果が期待される．商用データベースシステムと代表的なデータベースベンチマークである TPC-C を用いた実験により，30 秒から 100 秒程度のサービス停止時間のオーバヘッドのもとで，二次系ディスクストレージの消費電力のうち 80〜85%を削減可能であることを示した．10 秒以下の系切換を求める超高可用システムへの適用は難しいが，上記のオーバヘッドを許容可能なディザスタリカバリシステムなど広範なレプリケーションシステムへの適用が期待される．今後，更新情報のコンパクト化の有用性について，TPC-C 以外の多様なアプリケーションモデルについて検証を深めたい．

本論文では最近のストレージシステムにおける主要コンポーネントであるディスクドライブのみに着目したが，今後は，本アプローチを拡張し，フラッシュデバイス等の新たな記録媒体について検討を深めるほか，RAID コントローラやキャッシュメモリなどを考慮したシステム全体の解析を行い，有効性を検証したい．

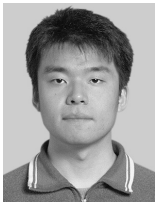
謝辞 本研究の一部は，文部科学省次世代 IT 基盤構築のための研究開発「非順序型実行原理に基づく超高性能データベースエンジンの開発」の助成により行われた．協力企業である株式会社日立製作所より有益なコメントを頂戴した．感謝する次第である．

文　献

[1] B. Rudolph, “Storage in an age of inconvenient truths,” Storage Network World Spring 2007, 2007.
[2] S.W. Worth, “Green storage I: Economics, environment, energy, and engineering,” SNIA Education, 2008.
[3] T.C. 9.9, Datacom Equipment Power Trends and Cooling Applications, ASHRAE, 2005.
[4] P.B. Chu and E. Riedel, “Green storage II: Metrics and measurement,” SNIA Education, 2008.
[5] IDC, “The diverse and exploding digital universe: An updated forecast of worldwide information growth through 2011,” An IDC White Paper sponsored by EMC.
[6] Q. Zhu, Z. Chen, L. Tan, Y. Zhou, K. Keeton, and J. Wikes, “Hibernator: Helping disk arrays sleep through the winter,” Proc. ACM Symp. on Operating Syst. Principles, pp.177–190, 2004.
[7] 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.
[8] S. Kleiman, “Trends in managing data at the petabyte scale,” Invited talk, USENIX Conf. on File and Storage Tech., 2007.
[9] Eagle Rock Alliance, “Online survey results: 2001 cost of downtime, contingency planning research,” 1996.
[10] 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.
[11] U.S. SEC, “General rules and regulations promulgated under the securities exchange act of 1934,” 2005.
[12] British Standards Institution, “BS25999: Business continuity managemenet,” 2006.
[13] EMC Corp., “Symmetrix remote data facility product description guide,” 2000.
[14] Hitachi Ltd., “Hitachi relational database management system solutions for disaster recovery to support business continuity,” Review Special Issue, Hitachi Technology, 2004.
[15] M. Ji, A. Veitch, and J. Wikes, “Seneca: Remote mirroring done write,” Proc. USENIX Conf. on File and Storage Tech., pp.253–268, 2003.
[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] Y. Lu and G. Micheli, “Comparing system-level power management policies,” IEEE Des. Test Comput., vol.18, no.2, pp.10–19, 2001.
[19] T. Lahiri, A. Ganesh, R. Weiss, and A. Joshi, “Fast-start: Quick fault recovery in Oracle,” White paper, 2001.
[20] 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," 情処学論（データベース），vol.46, no.SIG 8(TOD 26), pp.130–147, 2005.
[21] H. Lu and K. LeeTan, "Batch query processing in shared-nothing multiprocessors," Proc. Int'l Conf. on Database Syst. for Advanced Applications, pp.238–245, 1995.
[22] M. Mehta, V. Soloviev, and D.J. DeWitt, "Batch scheduling in parallel database systems," Proc. IEEE Int'l Conf. on Data. Eng., pp.400–410, 1993.
[23] D. Li, J. Wang, and P. Varman, "Conserving energy in conventional disk based RAID systems," Proc. Int'l Workshop on Storage Network Arch. and Parallel I/Os, pp.65–72, 2005.
[24] E.V. Carrera, E. Pinheiro, and R. Bianchini, "Conserving disk energy in network servers," Proc. Int'l Conf. on Supercomputing, pp.86–97, 2003.
[25] X. Yao and J. Wang, "RIMAC: A novel redundancy-based hierarchical cache architecture for energy efficient, high performance storage system," Proc. EuroSys, pp.249–262, 2006.
[26] HGST Inc., "Quietly cool," White Paper, HGST, 2004.
[27] 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.
[28] 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.
[29] R.A. Golding, P. Bosch, C. Staelin, T. Sullivan, and J. Wilkes, "Idleness is not sloth," Proc. USENIX Tech. Conf., pp.201–212, 1995.
[30] D. Colarelli and D. Grunwald, "Massive arrays of idle disks for storage archive," Proc. ACM/IEEE Conf. on Supercomputing, pp.1–11, 2002.
[31] C. Weddle, M. Oldham, J. Qian, A.A. Wang, P. Reiher, and G. Kuenning, "PARAID: A gear-shifting power-aware RAID," Proc. USENIX Conf. on File and Storage Tech., pp.245–260, 2007.
[32] S. Gurumurthi, A. Sivasubramaniam, M. Kandemir, and H. Franke, "Reducing disk power consumption in servers with DRPM," Computer, vol.36, no.12, pp.59–66, 2003.
[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. Magn., vol.36, no.5, pp.2213–2215, 2000.
[34] K. Okada, N. Kojima, and K. Yamashita, "A novel drive architecture of HDD: Multimode hard disc drive," Proc. Int'l Conf. on Consumer Electronics, pp.2213–2215, 2000.
[35] A. Weissel, B. Beutel, and F. Bellosa, "Cooperative I/O – A novel I/O semantics for energy-aware applications," Proc. USENIX Symp. on Operating Syst. Design and Imple., pp.117–130, 2002.
[36] Y. Lu, L. Benini, and G. Micheli, "Power-aware operating systems for interactive systems," IEEE Trans. Very Large Scale Integr. (VLSI) Syst., vol.10, no.2, pp.119–134, 2002.
[37] T. Heath, E. Pinheiro, J. Hom, U. Kremer, and R. Bianchini, "Application transformations for energy and power-aware device management," Proc. Int'l Conf. on Parallel Arch. and Compilation Tech., pp.121–130, 2002.
[38] S.W. Son, M. Mandemir, and A. Choudhary, "Software-directed disk power management for scientific applications," Proc. IEEE Parallel and Distributed Processing Symp., p.4b, 2005.
[39] 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.
[40] S. Quinlan and S. Dorward, "Venti: A new approach to archival storage," Proc. USENIX Conf. on File and Storage Tech., pp.89–102, 2002.

（平成 21 年 6 月 5 日受付，10 月 5 日再受付）



合田　和生

平 12 東大・工・電気卒．平 17 同大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．同年，博士（情報理工学）．日本学術振興会特別研究員等を経て，現在，東大生産技術研究所特任助教．超高性能データベースエンジン，高機能ストレージシステムの研究に従事．情報処理学会，日本データベース学会，ACM，IEEE CS，USENIX 各会員．

喜連川　優　（正員：フェロー）

昭 53 東大・工・電子卒．昭 58 同大大学院工学系研究科情報工学専攻博士課程了．工博．同年同大生産技術研究所講師．現在，同教授．平 15 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，東京支部長，情報処理学会副会長及びフェロー，日本データベース学会理事．ACM SIGMOD Japan Chapter Chair，本会データ工学研究専門委員会委員長歴任．VLDB Trustee，IEEE ICDE，PAKDD，WAIM などステアリング委員，SNIA 日本支部顧問，文科省特定領域研究「情報爆発 IT 基盤」領域代表を務める．