デペンダブルストレージシステム
－ 自己再編成ストレージシステム －

喜連川優
情報理工学系研究科 電子情報学専攻

概要

ストレージシステムのデペンダビリティに対する期待はとりわけ，９・１１以降，大きい．業界によっては，法制度化により，システム構成への信頼性強化が強いられる状況にもなりつつある．また，e-business における予測不能な負荷変動への対応も強く求められている．

一方，ストレージネットワーキング技術が近年大きく進展しており，各種のデータ管理アプリケーションがストレージシステム内において，そのプロセッサやキャッシュメモリを利用して実装されるに至っており，当該技術を利用した新たなソリューションが模索されている．

本研究では，データベース管理において不可欠な機能であるデータベース再編成をサーバシステムからストレージシステムへ移譲すること提案し，実装した試作機を用いて当該アプローチの有効性を明らかにする．

1 はじめに

ストレージ技術の潮流は大きく変化しつつある．デバイス技術の進展によりプロセッサやメモリは高性能かつ低価格になり，多くのプロセッサ，広大なキャッシュ空間，高速な内部スイッチ，及び多数のディスクドライブが備えられた大規模ストレージ装置が登場している．また，ストレージネットワークの普及により，大規模ストレージ装置は複数のサーバから共有される傾向にあり，即ち，ストレージを中心としてその周辺にサーバが接続されるストレージ中心のシステム構成が広く採用されるようになりつつある．

豊富な計算機資源を有する大規模ストレージ装置を中心とするシステム構成においては，IO インテンシブなデータ処理を，サーバ上の計算機資源を用いて実行するのではなく，大規模ストレージ装置の有する計算機資源の一部を用いて実施することに，以下のような利点が見受けられる．

1. **サーバ-ストレージ間 IO 帯域におけるボトルネックの解消**：一般にストレージ装置は高い内部帯域を有するのに対し，サーバ-ストレージ間の IO 帯域は比較的低い．サーバ上で IO インテンシブなデータ処理を行う場合，当該 IO 帯域が飽和し，ボトルネックとなりやすい．IO インテンシブなデータ処理をストレージ装置内で実行することにより，このようなボトルネックを解消することができる．
2. **物理情報に基づく IO 最適化制御**：IO インテンシブな処理の高速化には物理情報を用いた IO 最適化制御が不可欠である．仮想化された記憶空間へアクセスするサーバ上のアプリケーションには，物理情報の取得が難しい．一方，ストレージ装置内では個々のドライブの物理情報を容易に取得することができるため，データ処理に係る IO をより最適化し，高いディスク並列度を得ることが可能である．
3. **ストレージ側資源スケジューリング**：ストレージは複数のサーバから共有されるため，ストレージ装置の資源スケジューリングは単一のサーバに閉じず，システム全体の最適化は難しい．むしろ，ストレージ装置においてデータ処理を実行することにより，システム全体を見渡したグローバルなスケジューリングがより容易となる．
4. **データとその処理機能の共有資源化**：データとその処理機能をストレージ装置に置き，共有資源と見なしシステム全体で共有することにより，個々のサーバにおけるデータ処理を軽減させ，システム設計を容易化し，牽いてはシステムの TCO 削減が期待される．

既にバックアップやスナップショットなどの多くの機能が，ストレージ装置の持つ豊富な計算

機資源を有効に活用してストレージ装置内で実行されるようになっており，商用的にも広く受け入れられている．

現状においては，これらのデータ管理機能は，従来サーバ上で行われていた低レベルのデータ処理がストレージ上で実行されているに過ぎないが，ストレージ装置の有する計算機資源はより拡大しており，今後ストレージシステムが，DBMS の一部機能等のより高度なデータ管理機能を取り込む可能性は十分にある．

以上の技術背景を鑑み，本研究では，オンラインデータベース再編成機能を有する高機能ディスクストレージである自己再編成ストレージ（SRS: Self-Reorganizing Storage）システムを提案し，その有効性を検証する．再編成はデータベース管理における重要な機能の一つであり，構造劣化（Structural Deterioration）と呼ばれるデータベースにおける物理格納構造の乱れを解消する機能であり，極めて IO インテンシブである特色を有する．また，高可用性の要請から再編成はオンラインで実行可能であることが必須となりつつある．

SRS におけるオンライン再編成方式自体は，既に商用の DBMS 等で採用されている分離再編成なる方式に基づいているが，本研究ではデータと再編成機能がストレージシステム内に共存することの利点を追求する．即ち，ストレージ装置が有する豊富な IO 帯域と高い IO 処理能力を活用し，並列パイプライン化データ処理，物理アドレスレベルの IO 最適化，並びに独自の高速ログ適用処理なる技術を導入することにより，再編成を高速に実施することを目指す．著者の知る限り，データベース再編成をストレージシステム上で実行することを提案し，その有効性を検証する研究はこれまで行われていない．

本報告の構成は以下の通りである． 2. では SRS におけるストレージ側分離データベース再編成方式を示し，さらに，データベース再編成とログ追い付きを高速化する特徴的な要素技術を紹介する．3. では商用 DBMS を対象とした SRS の試作機の実装を紹介し，試作機による TPC-H 及び TPC-C ベンチマークを用いた性能評価実験を紹介し，有効性を論じる．4. では本報告をまとめる．

2 自己再編成ストレージ

自己再編成ストレージ（SRS）の基本的なアイデアはデータベース再編成なるデータ管理機能をサーバからストレージ装置へ移譲することにある．本節ではストレージ側でオンラインデータベース再編成を行うためのストレージ側分離データベース再編成方式を説明するとともに，ストレージ装置の有する高い IO 帯域と IO 処理能力を有効活用するための実行方式を述べる．

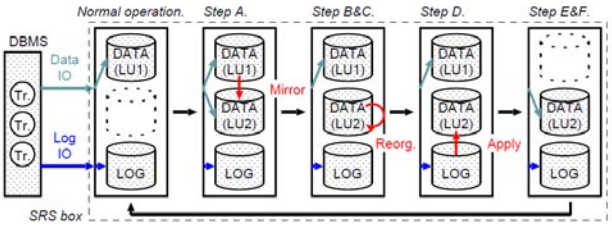

図 1 サーバ側分離データベース再編成

2.1 ストレージ側分離データベース再編成

SRS はストレージ装置の有するプロセッサ並びにキャッシュメモリを用いて，データベース再編成を実施する．その手順を図 1 に示す．

（A）先ず，データベース空間に対するミラー空間を動的に作成する．（B）次に，データベースサーバはデータベース空間の静止化を行い，ミラー対の分離を行い，複製元空間をマウントして，静止化を解除する．（C）ストレージ装置のプロセッサ上で実行されるデータベース再編成ソフトウェアは，複製先空間の再編成を実施し，（D）引き続いて，再編成された複製先空間をトランザクションの実行されている複製元空間に追い付かせるため，データベースのログ適用を行う．その後，（E）データベースサーバは再びデータベース空間の静止化を行い，再編成された複製先空間をマウントし，静止化を解除する．（F）最後に，複製元空間を解放する．

ストレージ側で実行する分離方式のデータベース再編成によって，データベース再編成をオンラインで実行することが可能となり，また，その際のトランザクション処理性能への副作用を最小限化することが期待される．即ち，一般にデータベース再編成は極めて IO インテンシブな処理であり，サーバ側でのオンライン再編成では，排他処理やサーバ-ストレージ間の IO 帯域がボトルネックとなるが，本提案方式では，ストレージ装置の多数のディスクドライブ，高い IO 帯域並びに IO 処理能力を活用することにより，この問題の解決を行う．

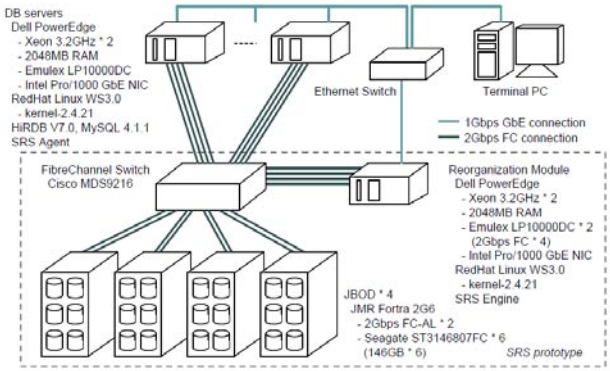

図 2 試作機と実験システム

上記の手順のうち，A., B., E., F. は DBMS の有する静止化機能及び，ストレージ装置が有する RAID 機能や仮想化機能，PiT コピー作成機能を用いて実現することができるため，本報告では詳細を述べない．一方，C.（データベース再編成）および D.（ログ追い付き）の手続きはストレージの新しい機能であり，実質的にこの手続きを高速化することが再編成の高速化につながることから，以降の節においてはその詳細に焦点を当てる．

2.2 並列パイプライン化データ処理と物理アドレスレベル IO 最適化

データベース再編成，並びにそれに続くログ追い付きはいずれも極めて IO インテンシブな処理であることから，ストレージ装置内での実装により，大幅な高速化が期待される．本節では，ストレージ装置内部の豊富な IO 帯域と高い IO 処理能力を効果的に活用するための実行方式として，並列パイプライン化データ処理，並びに物理アドレスレベル IO 最適化を述べる．

SRS ではデータベース再編成，並びにログ追い付きにおける処理スループットの高速化のため，データ処理を並列パイプライン化して実施する．例えば，所謂アンロード-ロード方式のデータベース再編成を行う場合，アンロード，整列，ロードは，可能な限りにおいて並列にパイプライン処理される．（この場合，整列ラン長は利用可能なメモリ長に制限される．）

加えて，一般にサーバ側の実装においては，データ処理は論理アドレス空間に対して実行されるが，この場合，IO は複数の仮想化層を経由するため，重積したオーバヘッドが無視できず，また，物理アドレスが隠蔽されていることから，スケジューリングなどによって十分に最適化を行うことはできない．対して，SRS においては，データベース再編成，並びにログ追い付きにおけるデータ処理を物理アドレスレベルで行うとともに，係る IO を同様に物理アドレスレベルでスケジューリングすることにより，IO スループットの飛躍的な向上が期待される．

2.3 高速ログ適用技術

前節では，ストレージ装置の高い内部 IO 帯域と IO 処理能力を効果的に活用するため実行方式を示したが，本節では別のアプローチとして，ログ追い付きにおけるログ系列の論理的な最適化による高速化を示す．前節の提案とは相補的に機能するものである．

一般に，データベースのロールフォワードではログ中の各エントリは論理時刻（LSN）順に適用される．対して，SRS においては，分離再編成におけるログ追い付きが一定のログ系列を一括して適用することに着目し，ログ系列をウィンドウバッファ内で論理的に最適化することを提案する．具体的には，ログ畳み込み，並びにログ整列なる 2 つの方式によってログ適用の最適化を行う．

ログ畳み込みは，データベース中の同一のレコードに対して更新を行う複数のログエントリを集約する．これにより，ログ追い付きによって適用するログエントリ数を削減することが期待される．ログ整列は，LSN 順ではなく，適用先のデータベースレコードの物理アドレス順にログエントリを適用する．これにより，一般にはランダムアクセスとなるログ適用の IO 系列を，よりシーケンシャル化することが可能となる．一般に，データベースのアクセスは局所性を有していることから，両方式を併用することにより，ログ追い付きの飛躍的な高速化が期待される．

3 試作機の実装と性能評価

3.1 試作機の実装

SRS の有効性を検証するため，図 2 に示すように，実装プラットフォームとしては，PC サーバ，ファイバチャネルスイッチ，ファイバチャネル JBOD を用い，SRS の試作機を実装した．PC サーバ上でサーバ側分離データベース再編成を実行するソフトウェアを実行する．ソフトウェアは，商用 DBMS である HiRDB，及びオープンソース DBMS である MySQL 双方に対応している．更に，

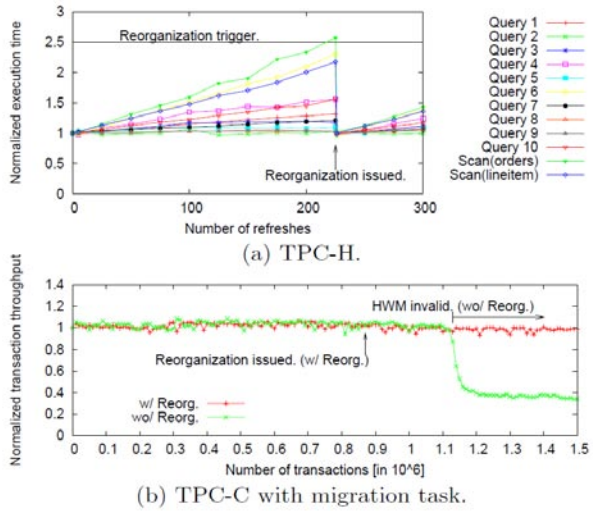

図 3 データベース再編成のケーススタディ

試作機には，データベースサーバと，管理用 PC 端末を接続し，実験システムとして運用する．

3.2 構造劣化とデータベース再編成

本節では，ケーススタディとして代表的なデータベースベンチマークである TPC-H 並びに TPC-C を用いて具体的な構造劣化を示すと共に，SRS によるデータベース再編成によって構造が回復する例を示す．

図 3（a）には，TPC-H において，データウェアハウスの更新に対する，各問合せの実行時間の変化を示す．更新によって一部の問合せの実行時間が徐々に悪化しているが，SRS は閾値によってデータベース再編成を駆動し，これにより構造が回復し，問合せ実行時間が改善している．また，図 3（b）には，TPC-C において，オンライントランザクション実行中にスループットが急激に低下するような構造劣化に対し，予め閾値を設けたデータベース再編成によって性能低下を回避する例を示す．上記，２つのケーススタディにより，SRS によるデータベース再編成によって，データウェアハウス，並びにオンライントランザクション処理に見られる構造劣化に対して，構造の回復を行い，性能の改善，もしくは急激な性能の低下を回避することが可能であることが分かる．

3.3 データベース再編成の性能評価

本節では，TPC-H 並びに TPC-C のデータセットに対して，従来的なサーバ側実装のデータベース再編成と，SRS におけるデータベース再編成の性能比較を行う．

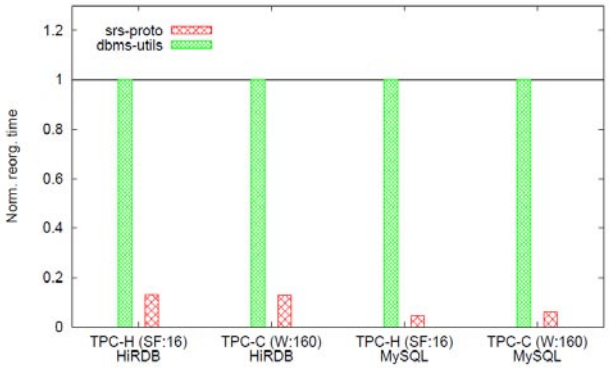

図 4 データベース再編成の性能比較

図 4 に計測結果を示す．HiRDB 及び MySQL 双方の DBMS において，SRS によるデータベース再編成（凡例：srs-proto）は，従来的なサーバ側の実装である再編成ユーティリティ（凡例：dbms-utils）と比較して，それぞれ約 13%，約 5%の実行時間でデータベース再編成を完了していることが分かる．このことから，今日一般的に用いられている再編成ユーティリティと比較して，SRS のデータベース再編成により大幅な性能改善が可能であることが分かる．

4 まとめと今後の課題

本研究ではデータベース再編成機能を有する高機能ディスクアレイである自己再編成ストレージ（SRS）を提案した．SRS はデータベース再編成をその内部で実施することにより，データベースの構造の効率性を保つことが可能である．ディスクストレージ内の高い IO 処理能力と豊富な IO 帯域を有効に利用するため，並列パイプライン化データ処理，物理アドレスレベル IO 最適化，高速ログ適用処理を実施する．商用 DBMS 並びにオープンソース DBMS を対象とした試作機を実装し，性能評価実験を行った．実験の結果，提案的なサーバ側の実装による手法と比較して，１桁高速なデータベース再編成が実現されたことが示された．

自己再編成ストレージは，構造劣化の管理をサーバからストレージに移譲しようとするアプローチの提案である．より詳細なストレージとサーバの新しい役割分担の有効性について，検討を進め，ストレージシステムのデペンダビリティ向上に活かしたい．