6L-3 Storage Fusion: 自己再編成ストレージシステムの高度化に関する一考察

星野喬† 合田和生‡ 喜連川優‡

† 東京大学 大学院情報理工学系研究科 ‡ 東京大学 生産技術研究所

概 要

更新によってデータベースに生じた構造劣化を除去し，性能を改善するデータベース再編成は，データベースの性能管理に不可欠な機能である．本論文では，一般にデータベース再編成はその処理がデータインテンシブであり，DBMS で実行される問合せ処理などの高レベル処理とは本来的に分離が可能であるという特性に着目し，データベースの構造劣化の管理をストレージに移管することの有益性を議論する．著者らの開発した自己再編成ストレージシステムと称する自律的な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の試作機を示し，その高度化手法に関して考察する．

1 はじめに

巨大ストレージシステムの登場とそれによるストレージコンソリデーションの進展によって [1,2]，サーバとストレージの間には，新たな機能分割の可能性が与えられるようになった．即ち，従来型の IT システムでは，全てのソフトウェアは原則的にサーバのプロセッサ上で実行されていたのに対し，今日の IT システムにおいては，ソフトウェアの一部をストレージシステムのプロセッサ上で実行することにより，サーバとストレージが役割分担を行うことが現実的となりつつある．このような新しい機能分割を検討することは，データ管理を担うシステムアーキテクチャの重要な研究課題である．

サーバとストレージの間の機能分割に関する学術的な研究として，著者らは，データベースの構造劣化の管理をストレージに移管することの有益性を議論してきた [4]．構造劣化とは，データベースの更新によって生じた格納構造の変化が問合せ処理性能を低下させる現象であり，当該劣化現象の管理はデータベースシステムにおける性能管理上，本質的な課題である．データベース管理者はデータベースの構造劣化の具合を把握し，必要に応じてデータベースの再編成を行うことにより，データベースの性能を維持する役目を負っている．しかし，一般に，このような構造劣化の管理は容易な業務ではなく，しばしばデータベース管理者の悩みの種と称される．データベース再編成は極めてデータインテンシブな処理である特色を有する一方，DBMS で実行される問合せ処理やトランザクション処理などの高レベル処理とは本来的に分離が可能である．データベース再編成を自律的にストレージシステム内で実行できるようにすることにより，再編成の効率的な実行が実現され，方やデータベースサーバはもはや構造劣化を意識する必要はなくなり，システム全体で管理業務が大幅に軽減されることが期待される．ストレージレベルの実装は新しく有力なアプローチといえる．

本論文では，著者らが開発してきた自己再編成ストレージシステムと称する自律的なデータベース再編成を可能とする高機能ストレージシステムについて，その高度化を議論する．即ち，当該システムにおいて，より高い再編成スループットを達成することを目的として，再編成処理を複数の演算ノードである制御モジュールに分散させる並列再編成機構を議論し，試作機を用いた性能評価を示す．

2 データベース再編成処理の並列化

論文 [4] では，システムアーキテクチャとして対象型マルチプロセッサ (SMP) 方式を採用してきた．即ち，同一バス上にプロセッサとメモリを配置し，全てのプロセッサからバス上のメモリ及びディスク等のデバイスが対称にアクセス可能であった．しかし，このような SMP アーキテクチャにおいては，再編成に代表されるデータインテンシブな処理において，十分なディスクアクセスの並列度を得ようとする場合，バス性能及びプロセッサ演算能力が飽和する可能性がある．SMP アーキテクチャにおいて，バス帯域を拡張し，プロセッサ数を増加させることは，価格性能比の面からも有効でない．このため，当該ストレージシステムにおける再編成の更なる高スループット化を目指す上で，障害となる可能性がある．

本論文では，図 1 に示すように，自己再編成ストレージの内部において，複数の制御モジュールに渡って再編成プロセスを分散させる再編成の並列化方式を示す．再編成処理を複数の制御モジュールに分散可能とすることにより，バス帯域及びプロセッサ演算能力の拡張が可能となり，更なる再編成の高スループット化が期待される．ここに，制御モジュールはプロセッサに加えて，ローカルメモリとしての主記憶を有する．ディ

Storage Fusion: A Study on Further Speed-up of Self-Reorganizing Storage System

HOSHINO Takashi†and GODA Kazuo‡and KITSUREGAWA Masaru‡

†Graduate School of Information Science and Technology, The University of Tokyo ‡Institute of Industrial Science, The University of Tokyo

{hoshino,kgoda,kitsure}@tkl.iis.u-tokyo.ac.jp

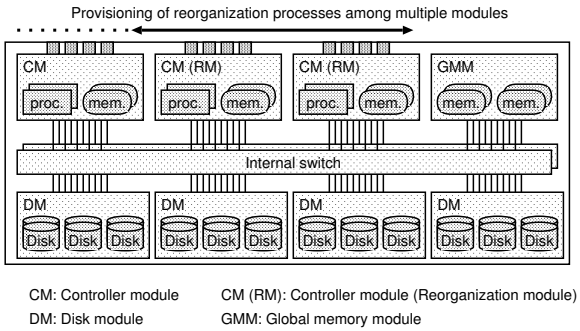


図 1: システムアーキテクチャ.

スクストレージ内で複数のモジュールに処理を分散させる内部分散化アーキテクチャは，既に商用ストレージシステムにおいても採用されており，本方式は産業上も，実現可能なものと考えられる．

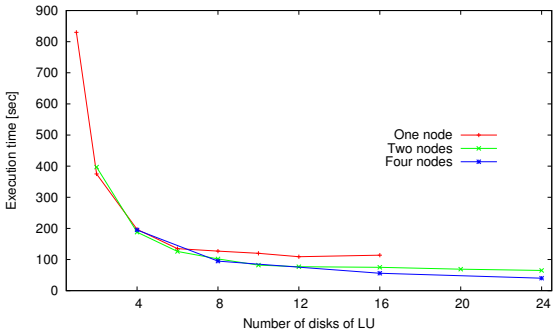
3 試作機による性能評価

著者らは，自己再編成ストレージの試作機において，再編成処理を複数の制御モジュールを用いて実行する改造を行った．紙面の制約の都合上，詳細は別稿に譲る．また，当該並列再編成処理の有効性を検証するために，商用 DBMS である HiRDB [3] 上で代表的なデータベースベンチマークである TPC-H をアプリケーションとして用い，データベース再編成の性能測定を行なった．

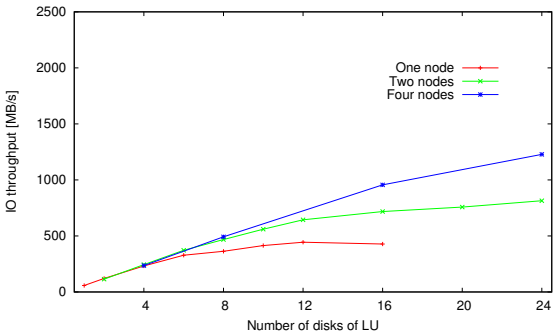
TPC-H におけるスケールファクタを 16 とした初期データに対して，RF1 及び RF2 による更新を 10 回行ったものを対象に，目標充填率を 100%としたデータベース再編成を実施し，ボリュームを構成するディスク台数を 1 から 48 まで変化させ，データベース再編成の実行時間，IO スループットを測定した．この際，再編成ソフトウェアが稼働する制御モジュールの台数を 1 から 4 まで変化させて計測した．ボリューム内のディスクは常に等しい数で制御モジュールに割り当てられるものとした．

ボリューム内でデータベース再編成を実施した場合の計測結果を図 2 に示す．ディスクの台数が 6 台以下の場合，ノード数に関係なく，同様の測定結果が得られており，大きな相違はないと言える．1 ノードの場合は，ディスク数が 8 より多い場合，スループットの伸びが鈍化しており，これ以上の実行時間の改善は期待できない．一方，4 ノードの場合では，ディスク台数が 9 台以上に於いても実行時間，IO スループットが改善しており，最大でディスク台数 24 台までスケールすることができ，1160MB/s の IO スループットを達成している．

本実験結果から，データデータベース再編成に於いて複数の再編成モジュールに分散可能な再編成処理が有効であることが分かる．



(a) Execution time.



(b) IO throughput.

図 2: ディスク並列度とデータベース再編成性能の関係．(ボリューム内再編成).

4 おわりに

本論文では，著者らの開発した自己再編成ストレージシステムと称する自律的な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の高度化手法として，複数ノードを用いた並列再編成手法に関して議論し，試作機を用いた実験によりその有効性を示した．

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．

参考文献

[1] EMC Corp. EMC Symmetrix DMX Series. White Paper, 2004.

[2] N. Takahashi and H. Yoshida. Hitachi TagmaStore Universal Storage Platform: Virtualization without Limits. White Paper, Hitachi Ltd., 2004.

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

[4] 合田, 喜連川. 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, 46(SIG 8(TOD 26)):130–147, 2005.