論 文 の 内 容 の 要 旨

論文題目　　大規模並列関係データベース処理における高速化技法に関する研究

氏　名　　中野　美由紀

二十一世紀に入り、情報化社会の発展はますます加速し、インターネットの急速な普及に伴い、ワールド・ワイド・ウェブ等に見られるように、個人が世界に向けて情報発信が可能となった時代を迎えている。それに伴い、様々な人類の活動の記録は、電子的データとして日々蓄積され、その容量は爆発的に増大している。関係データベースシステムは計算機の出現により始まった情報化社会の発展に伴い、銀行のオンライン処理、航空機、電車などの座席予約システム、製造業、流通業の在庫管理システム、企業内人事管理システムなど、ビジネスにおけるあらゆる側面において必要不可欠なシステムとして利用されるようになった。さらに、インターネット上を多量の情報が流通する現在、電子商取引、各種のウェブサービス、また、ウェブ上にある多種多様な大容量コンテンツ管理およびその情報解析等において、関係データベースシステムは必要不可欠な基盤技術（ミドルウェア）と認識されている。

一方、技術の進歩に伴い、共有メモリ計算機、分散メモリ計算機、分散共有メモリ計算機などアーキテクチャの異なる並列計算機が提案され、現在では１０００台規模を超えるクラスタシステムが大規模データセンタ、ウェブ情報の検索エンジンなどで広く利用され

ている。また、ピア・ツー・ピア（Peer to Peer）システムなど、ネットワークを介した個々のパソコン、ワークステーションの計算資源を他のユーザなどにも供することで、広く分散処理を目指したシステム・アーキテクチャも実際に運用されつつある。結果、オラクルのOracle10i、IBMのDB2、マイクロソフトのSQL Serverなど広く用いられる商用関係データベースシステムも、急増するデータ量、計算機技術の進歩によるアーキテクチャの進化に伴い、常に処理の高速化、性能向上を求められ、一年ごとにシステムが改版されていくのが実情である。

このように、関係データベースシステムは、社会基盤の不可欠な要素の一つとして、多様に変化する利用者側の高性能、高機能への要求に答えると共に、新たに現れる異なるアーキテクチャ上への効率のよい実装が常に望まれてきた。しかしながら、関係データベースシステムの並列処理技法にはいまだ多くの課題が残っており、スケーラブルなシステム拡張を容易とする高性能、高機能並列データベースシステムの研究が急務となっている。

本研究は、大規模データを扱う関係データベースシステムの並列処理方式に関し、異なる並列計算機環境上において、ストレージの入出力制御方式、データベースシステム内のバッファ管理方式、関係データベース演算処理の並列化方式、関係データベース問合せ処理方式の観点から、高性能化、高機能化技法の開発、評価を行っている。

ストレージのアクセス手法およびシステム上のバッファ管理機構は関係データベース処理性能を向上させる重要な要素である。汎用OS上で提供される入出力処理機構とマルチスレッドを単純に利用してストレージアクセスを行っていては、関係データベース上の大規模データを効率良く処理することはできない。そこで、関係データベースシステム上で必要な入出力機構、フィルタリング機構を抽出し、従来のOSとは異なるモジュール分割を行うことで関係データベースシステムに適合した新たなシステム・アーキテクチャ「機能ディスクシステム」の提案を行った。さらに、試作機を実際に開発し、本研究で提案した関係データベースシステムに適合した入出力ドライバ、入出力ライブラリおよび共有メモリ管理機構ライブラリを既存の OS9 をベースとして新たに構築、試作機上に実装した。この試作機上において、QUEL をベースとする関係データベース問合せシステムおよび「機能ディスクシステム」のハッシュ機構と利用した並列問合せ処理システムを構築し、データベースベンチマーク(Wisconsin Benchmark)の性能測定を行った。この結果、従来の関係データベースシステムと比較し、百倍程度の高速化を達成したのみならず、Wisconsin 大学の

Gammaプロジェクトの結果と比較し、システム構成としては小規模ながらも、関係データベースシステムに特化した実装を行うことで、数十倍の性能向上が得られることを示した。

共有メモリ計算機は、近年では数百GBにおよぶ共有メモリを搭載したサーバが出現するに至っている。共有メモリ計算機では、いったん主記憶上にデータがロードされれば、並列処理は容易に実現できる。しかしながら、これらの大容量の主記憶を利用する場合、必要なデータをストレージからロードするコストは非常に重く、バッファ管理および入出力制御は単純な方式では性能向上が得られない。本研究は、ハッシュ関数を用いた並列問合せ処理の実装をめざし、ハッシュ結合演算処理の実装方式について、入出力バッファの柔軟な切り分けと共有メモリを有効利用について検討を行った。実際に、商用の共有メモリ計算機(Sequent)上にて、並列処理効果のハッシュ関数を用いた並列問合せ処理を実装し、その結果から提案する方式が高い並列処理効果を得ることができることを示した。また、ストレージからのデータロードと主記憶上での演算処理を重ね合わせて処理することにより、実装した方式がストレージの入出力転送幅を十分に利用可能であることを示した。

分散メモリ計算機はノード、ストレージ等の資源の追加することで、現状のデータの爆発的増加に対応したシステム拡張が容易である。また、共有メモリ計算機と異なり、高速なバスを必要としないため、コストパフォーマンスも良い。一方、それぞれのノードが個々に処理を行うため、処理の同期、負荷分散などを考慮しなくてはならない。分散メモリ計算機環境における多重並列結合演算処理方式の最適化方式は、主記憶、ネットワーク転送の利用を個別に考察したものはあるが、システム全体での計算機資源の利用均衡化という観点の研究は少ない。分散メモリ計算機では、あるノードで過剰な処理負荷が生じると、それがボトルネックとなり、全体の処理性能が低下する。そこで、分散共有メモリ計算機において主要な資源、ネットワーク転送、CPU処理、入出力転送のコストを用い、資源消費が均衡するような部分木を生成、候補木の集合とし、それらの部分木の組み合わせを基にdynamic programming法を用いて最小の処理コストとなる最終実行木を生成する方式を提案した。シミュレータを生成し、提案する最適化技法により生成される実行木のコストが従来の資源消費について考慮しない手法と比較して、品質に関しても、探索空間に関しても、十分によい結果木を導出可能であることを示した。

次に、分散メモリ計算機のシステム拡張性の利点および共有メモリ計算機の実装の容易性を併せ持つ分散共有メモリ計算機上において、分散共有メモリのアクセス特性に合わせ

てデータベースシステムのアクセス局所性を考慮した並列結合演算処理方式を提案した。分散共有メモリのアクセス特性に合わせてデータベースシステムのアクセス局所性を考慮した並列結合演算処理方式を提案し、商用分散共有メモリ計算機上に実装し、キャッシュすることにより生じるデータ参照局所性を利用することにより、一般にデータ局所性がほとんどないと言われる大規模意思決定支援システム上での問合せ処理などに関し、提案する方式が有効であることを示した。さらに、分散共有メモリ計算機のキャッシュコヒーレンシ機構に関するメモリアクセスコスト等について、実機上において詳細なデータをとり、これに基づき、シミュレーションを作成し、提案する方式が、キャッシュサイズが相対的に小さくなるようなノード台数が大幅に拡張された場合にも有効であることを確認した。また、あらかじめキャッシュ上の参照される頻度の高いデータを複製することで、データのアクセスに偏りが生じた場合にも、負荷分散が可能であることを示した。

本研究では並列関係データベース演算として、最も処理負荷の高い結合演算を対象として議論を進めてきた。結合演算の並列化アルゴリズムとしては、Grace ハッシュアルゴリズムを代表とするハッシュに基づく結合演算処理の性能がよい。しかしながら、電話帳の名前の分布などでも知られているように実世界のデータは偏っており、データの偏りによっては、均等に分割できる適当なハッシュ関数がない場合も多い。そこで、ハッシュ関数を用いた並列処理技法に従来のネストループ処理方式を組み合わせ、 GN Hash 方式を提案した。一般にネストループ方式は、内側のリレーションを複数回読み出すため処理コストは高いが、その処理コストはデータの偏りには依存しない。一方、ハッシュ結合演算では、ハッシュ関数適用後のそれぞれのクラスタのデータ分布が一様であるという保証はなく、データの偏りが高い場合にはクラスタの再分割を繰り返し行う必要がる。また、データ分布が一様であったとしても、外側のリレーションが主記憶の数倍程度の大きさであれば、ネストスープ方式の処理コストがハッシュ結合演算の処理コストよりも小さくなる。そこで、データの偏りによりクラスタの再分割が起こった場合には、ネストループ方式をそのクラスタに適応することで、繰返されるクラスタの再分割を防ぐことで、負荷の偏りにもロバストとなる。ハッシュ結合演算処理およびネストループ処理の詳細なコスト式を導出し、シミュレーションによる詳細な解析を行った。その結果、他の結合演算処理方式と比較し、GN ハッシュ方式がデータの偏りが非常に高い場合にも性能劣化が少ないことを確認した。