

4L-10

Symmetry S81 における結合演算の並列処理の性能評価

津高 新一郎 中野 美由紀 喜連川 優 高木 幹雄

東京大学 生産技術研究所

1 はじめに

関係データベース処理の中で、結合演算は、選択演算などの他の関係演算に比べて処理負荷が重いことは良く知られており、その処理負荷を軽減すべく今まで種々の結合演算処理方式が提案されてきた。なかでもハッシュ操作に基づく結合演算処理は、従来のソート処理に比べて高い性能が得られ、数々の研究成果が発表されている[1, 2, 3, 4]。

一方、データベースマシンの分野では、処理性能の向上を目指し、並列処理技法を取り入れたアーキテクチャが多数考案されている。我々は、共有メモリマルチプロセッサマシン上での関係演算処理方式として、最も処理負荷の重い結合演算の実装について検討している。3.2 MBの共有メモリ、18台のプロセッサ、8台のディスクから構成されるSequent社のSymmetryを用い、この上に並列処理技法を取り入れた結合演算処理方式を実装した[5, 6]。本稿では、その実装方式と性能評価について述べる。

2 実装の特徴

当研究室のSymmetry S81は、3.2 MBの共有メモリ、18台のプロセッサ(i386、16 MHz) 8台のディスクから構成される。8台のディスクは各々別のチャンネルに接続され、並列にアクセスすることが可能である。

実装する結合演算処理方式としては、Grace Hash方式を採用する。これは大規模なリレーション同士の結合演算に適したアルゴリズムであり、ハッシュ操作に基づいて演算処理対象を独立した複数の空間に分割し、結合処理時のデータ検索空間を狭めることで、演算処理の高速化を図る、というのがその基本的な考えとなっている(図1、2、3)。

Grace Hash方式を本マシン上に実装する際には、共有メモリ上のデータに対するマルチプロセッサによる並列処理効果を考慮すると共に、入出力処理に対する並列処理方式についても留意しなければならない。そこで、今回の実装方式の特徴を以下に示す。

プロセッサ割り当て プロセッサを2種類、すなわちディスクの入出力アクセス処理(図1、2、3で濃い矢印)を行なうプロセッサ群(入出力プロセッサ)と、その他のメモリ上での比較、転送などのデータ処理(図1、2、3で薄い矢印)を行なうプロセッサ群(ジョインプロセッサ)に分離する。入出力プロセッサは各ディスクに1台ずつ与えられ、入出力の集中管理を行なう。ジョインプロセッサと入出力プロセッサはリード/ライトキューによってデータの通信を行なう。これらによりディスクのノンストップ化と入出力処理とオンメモリ処理の両方における高度な並列化を実現する(図4)。

ファイル構成 水平分割によりリレーションを複数台のディスク上へ分割して配置し(ストライピング)、並列にアクセスを行なって入出力を高速化する。また、入出力コストを詳細に評価するため、ロードバースを使用した入出力を採用する。

3 性能評価

タプル長208バイト、結合属性長4バイト、属性値がユニークで順番がランダムであるようなリレーションに対し、64kバイトのページ256個からなる16Mバイトのメインメモリをステージングバッファとして用いて、ディスク台数、プロ

セッサ数、リレーションサイズをさまざまに変化させて結合演算を行ない、そのコストの測定を行なった。

3.1 リレーションの大きさと総コスト

まず、ディスクが8台のとき、10万件から100万件の大きさを持つリレーションに対し結合演算を行ない、そのコストを測定した。そのコストと、入出力にかかるコストを図5に示す。処理にかかる総コストはリレーションの件数にほぼ比例して増加し、また、そのうちのほとんどを入出力コストが占めており、オーバーラップ処理がほぼ完全に行なわれていることから、Symmetry上で我々の提案したGrace Hash実装方式が有効であることが確認できる。

3.2 ディスクの台数効果

次に、ディスクの数を1台から8台まで変化した場合に、10万件のリレーションの結合演算を行ない、その性能比を求めた(図6)。ディスク4台まではほぼ線形に性能向上が見られるが、それ以上の台数では急激に性能が低下し、ディスク8台では6程度である。その主な原因としては、入出力の単位を64kバイトと比較的大きくとってあるため、ディスク間でのリレーションの分割が不均等になるという点や、ディスク1台当たりのメモリが少なくなり、入出力バッファを圧迫するという点、また、ディスクの台数の増加により入出力時間にばらつきが生じ、一番遅いディスクがボトルネックとなってコストが上昇することなどが考えられる。

3.3 ジョインプロセッサの影響

総コストはディスクの台数、すなわち入出力プロセッサの数だけでなく、ジョインプロセッサの数にも依存する。10万件のリレーションを用い、ディスクの数とジョインプロセッサの数を変えて結合演算処理を行ない、そのコストを測定した(図7)。ディスクが8台のときはジョインプロセッサが4台以上、4台のときは2台以上でコストがほぼ一定となっている。ディスクが2台、1台の時はコストはジョインプロセッサの数にほとんど依存しない。このことから、ディスクの性能を十分に生かすためにはディスク(入出力プロセッサ)1台あたりおよそ0.5台のジョインプロセッサが必要であることがわかる。

4 おわりに

Grace Hash方式のSymmetry上における実装方法を述べるとともにその性能の評価を行ない、ハッシュを用いた処理方式が有効であることが確認された。関係データベース演算の中でも最も負荷の重い結合演算に対して、ディスクの台数効果が明確に認められ、共有メモリ型マルチプロセッサアーキテクチャが関係データベースシステムを構築するうえで極めて有効であることが確認された。また入出力を管理するプロセッサとその他の処理をするプロセッサの数をさまざまに変化させて測定し、最大の効率を与えるためのそれらの適正な比を求めることが出来た。

現在、Symmetry上でDynamic Grace Hashの実装が進んでいる。これは、ステージングバッファをさらに有効に利用することによって入出力コストを低下させることを目的とした、Grace Hashアルゴリズムの改良版である。また、類似のアルゴリズムとして、Hybrid Hashアルゴリズムがある。これらの実装を完成させ、リレーションのサイズによってどのようなアルゴリズムが最適なのか、またその時のプロセッサの割

Performance Evaluation of Parallel Processing of Join Operation on Symmetry S81

S.Tsudaka, M.Nakano, M.Kitsuregawa, M.Takagi
Institute of Industrial Science, University of Tokyo

り当てはどうあるべきか、といったことについて今後研究を進めていく予定である。

参考文献

- [1] 喜連川 優, 他. 「動的処理バケット選択手法に基づくハッシュ結合処理方式とその性能評価」, 情報処理学会論文誌第30巻, 1989.
- [2] M.Kitsuregawa, et al. "The Effect of Bucket Size Tuning in the Dynamic Hybrid GRACE Hash Join Method", VLDB 89, 1989.
- [3] D.J.Dewitt, R.Gerber. "Multiprocessor Hash-Based Join Algorithms", VLDB 85, 1985.
- [4] L.D.Shapiro. "Join Processing in Database Systems With Large Memories", ACM TODS, Vol.11, No.3, 1986.
- [5] 津高, 中野, 喜連川, 高木. 「Symmetry S81におけるGRACE HASH方式の実装と評価」, 情報処理学会第40回全国大会, 1990.
- [6] 津高, 中野, 喜連川, 高木. 「Symmetry S81における結合演算の並列処理に対する考察」, 情報処理学会第41回全国大会, 1990.

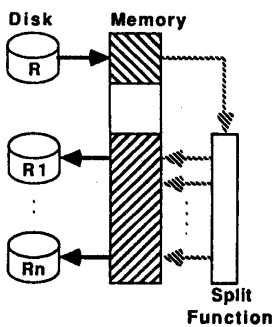


図1.Split Phase

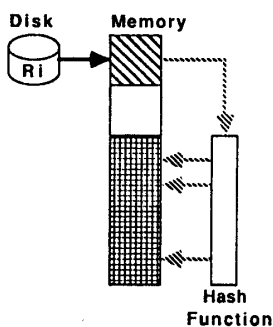


図2.Build Phase

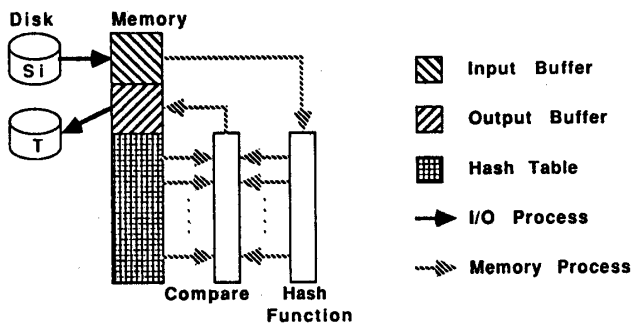


図3.Probe Phase

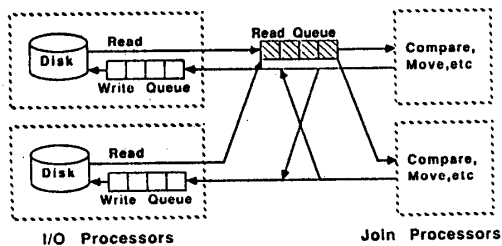


図4.

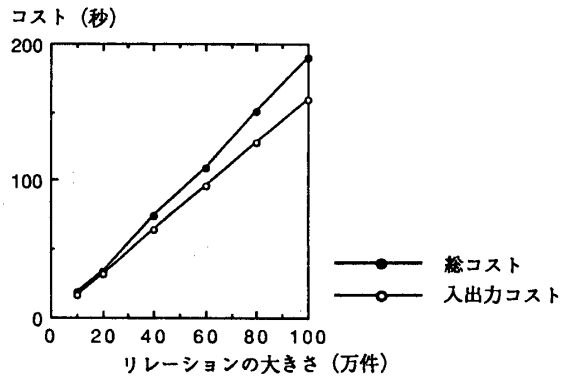


図5. 測定結果1

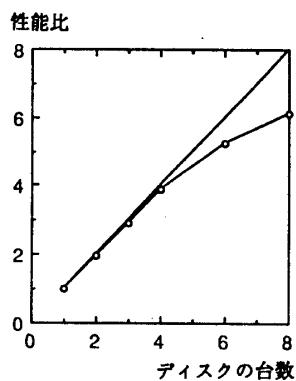


図6. 測定結果2

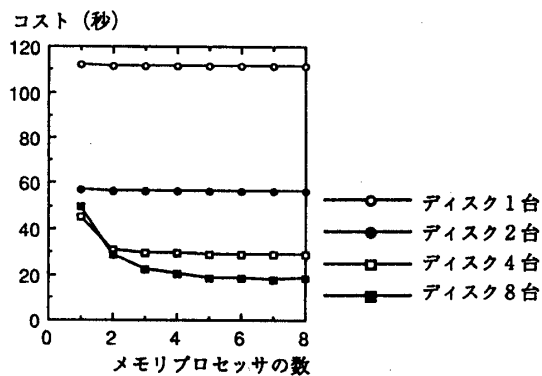


図7. 測定結果3