トランスピュータを用いた並列データベースマシンにおける結合演算の性能評価

5R−2

赤星 直輝　野口 泰生　萩原つね子　武 理一郎
(株) 富士通研究所

1　はじめに

近年、関係データベースの処理を高速化するための並列データベースマシンの研究が行なわれている[1]。我々は、トランスピュータを用いて並列データベースマシンを構成し、各種の並列アルゴリズムの研究をおこなっている。

我々のデータベースマシンは、Shared-Nothingのアーキテクチャを持つ。プロセッシングエレメント(PE)と、多段結合ネットワークは、トランスピュータにより構成されている。ネットワークの特徴としては、全対全通信の機能を持っている[2]。これまで、我々のマシンは、PEにディスクを接続しておらず、大規模なリレーションを処理することができなかった。今回、PEにディスクを接続して結合演算の実装と性能評価を行なったので、これについて報告する。

2　トランスピュータを用いた並列データベースマシンの構成

ここでは、我々が使用しているトランスピュータを用いた並列データベースマシンの構成について述べる。

2.1　トランスピュータの概要

トランスピュータは、Inmos社が開発したプロセッサであり、これを用いて並列処理のシステムが容易に構成できる。特徴としては、通信用のリンクを4本持っており、1次元や、2次元のアレイなど各種のトポロジを構成できる。また、トランスピュータを使用するための言語としては、処理の並列性の記述や、通信・同期の記述も可能なOccamや、Cを使用することができる。

2.2　並列データベースマシンの構成

我々の試作した並列データベースマシンは、PEと、それらを結合する多段結合ネットワークからなる。PEにはT800トランスピュータモジュールを32個、SCSIディスクコントロール用のT222トランスピュータモジュール16個と富士通製M2624SA 500MB 3.5インチSCSIハードディスクドライブ16台を使用した。ネットワークは1MBのメモリを持つT800トランスピュータモジュールを32個使用して、Binary n-cubeネットワークを構成している。この並列データベースマシンはホストのSUNワークステーションのバックエンドとして動作する。図1に計算機の構成を示す。

システム全体は、ソフトウェアによって、PE数を2から16まで変化させることができる。言語は、Occamを用いて、関係演算、各PE間や、PEとスイッチ間の通信、SCSIドライバといった処理を記述している。ネットワークの部分は、ハッシュ時に発生する全対全通信を効率よく処理するための機能を持つ[2]。

Performance Evaluation of Join Operation on Transputer Based Parallel Database Machine
Naoki Akaboshi, Yasuo Noguchi, Tsuneko Hagiwara, Riichiro Take
FUJITSU Laboratoires Ltd.

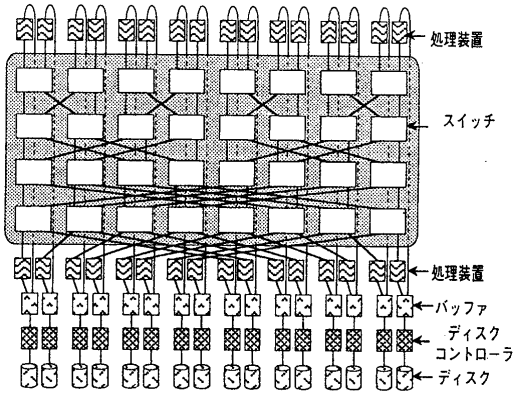

図1　並列データベースマシンの構成

3　並列データベースマシン上での結合演算

関係演算の中でも結合演算は、処理の重いものであり、その高速化に関して多くの研究が行なわれてきた。中でもハッシュ法は効率の良いものとして知られており[3,4]、今回の並列データベースマシンにもハッシュ法を用いた。

ハッシュ結合によって、同じハッシュ値を持つレコードの組み合わせを減らすことができる(図2)。特に並列データベースマシンでは、ハッシュした値によってどのPEで処理するかを決めておけば、他のPEと関係なく独立して結合処理することができる利点がある。

今回の実装では、結合演算の対象となるリレーションは全てのディスク上に分割して配置している(Full Declustering)。まず、各PEがディスクからリレーションRを読み込み、Rの各タ

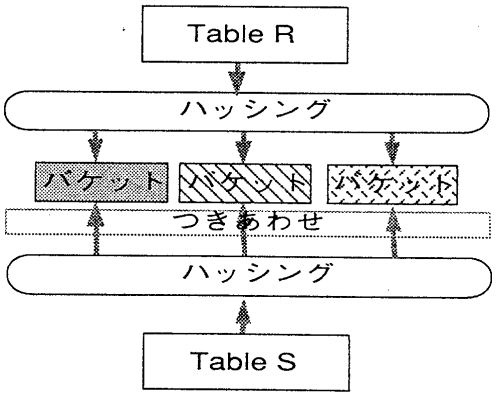

図2　ハッシュ結合

プル毎に結合演算を行なう属性の値によってハッシュ(Splitフェーズ)して、そのハッシュ値のタプルを処理するPEに対してタプルを送信する。このとき、ネットワークでは、全対全通信処理が発生する。各PEでは、受け取ったタプルのハッシュテーブルを作る(Buildフェーズ)。

次に、ディスクからリレーションSを読み込んで、各タプル毎に処理を行なうPEに送る。このときも、全対全通信が発生する。各PEは、Sのタプルを受け取ったら、ハッシュテーブル上のリレーションRとつきあわせて結合をおこない、結果をディスクに書き出す。

4 結合演算の性能評価

タプル長208バイト、結合属性4バイトのリレーションに対して、リレーションサイズやPEの数を変化させて結合演算の実験を行なった。

4.1 並列データベースマシンの基本性能

はじめに、トランスピュータを用いた並列データベースマシンの基本的な性能を表1に示す。疎結合並列データベースマシンでは、CPU、I/O、ネットワークの処理速度のうち、最も遅い部分が性能のボトルネックとなる。我々の並列データベースマシンでは、I/Oの処理速度が最も遅く、I/Oバウンドになっていることが分かる。ページサイズ16KBディスクの処理速度は、本来780KB/sec程度あるが、コマンド処理の回転待ちのため低下してしまっている。

ディスク転送速度	470 KB/sec
ディスクページサイズ	16 KB
全対全ネットワーク転送速度	780KB/sec
処理装置台数	2-16
処理装置1台あたりのメモリ	2 MB

表1 並列データベースマシンの基本性能

4.2 リレーションサイズによる影響

台数を16台とし、リレーションサイズを変化させたときの実験結果を図3に示す。リレーションサイズの増大に比例して、結合演算の処理時間は増加していることがわかる。これは、大量のタプルの全対全通信を効率よく処理しているため、ネットワークバウンドではなく、I/Oバウンドで処理がおこなわれるからである。また、トランスピュータ上では、これらの処理の記述をOccam言語により容易に実現することができる。100Kタプルの処理に23秒を要している。

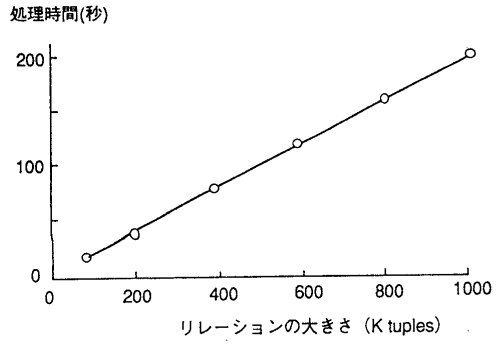

図3 リレーションサイズの影響

4.3 台数による影響

リレーションの大きさを100Kタプルと一定にした台数効果の実験結果を図4に示す。台数の向上と速度比が一定でない。これは、台数が増加するのにつれて、我々のシステムのネットワークインターフェースの処理が重くなるからである。

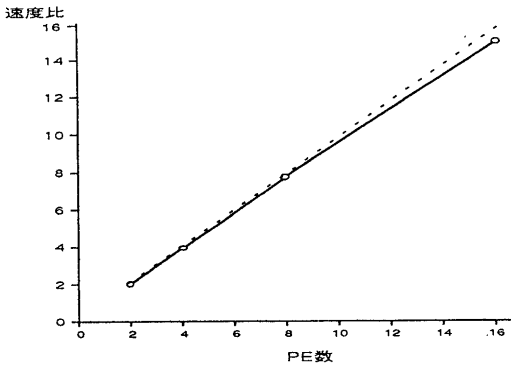

図4 台数効果

5 おわりに

本稿では、トランスピュータを用いた並列データベースマシン上での結合演算についての性能評価を行なった。並列データベースマシンでは、全対全通信を効率よく処理するネットワークを利用して、ハッシュ法を用いた並列結合演算の高速化が容易に行なえることを示した。また、トランスピュータ上では、Occam言語によりこれらの処理の記述を容易に実現できた。

今後は、各種の関係演算の効率よい処理についての検討や、ネットワークインターフェースの改良、I/O回りの最適化を行なう予定である。

参考文献

[1]D.J.DeWitt, S.Ghandeharizadeh, D.A.Schneider, A.Blicker, H. Hisao, and R.Rasmussen, "The Gamma Database Machine Project", *IEEE Trans. on Knowledge and Data Engineering*, Vol.2, No.1, 1990

[2]野口泰生、武理一郎、横田治夫、「分散制御型全対全通信結合網」、情報処理学会第42回全国大会、1991

[3]M.Kitsuregawa, H.Tanaka, and T.Motooka, "Application of hash to database machine and its architecture", *New Generation Computing*, 1(1), pp.66-74, 1983

[4]D.Shneider and D.J.DeWitt, "A Performance Evaluation of Four Parallel Join Algorithms in a Shared-Nothing Multiprocessor Environment",*Proc. of the 1989 SIGMOD Conf.*,pp.110-121,1989