並列計算機 AP1000DDV における多重結合演算の実装とその評価

7D－4

新谷隆彦, 中野美由紀, 喜連川 優
東京大学生産技術研究所

1 はじめに

近年、並列多重結合演算処理の研究では、ハッシュ結合演算を基に複数の結合演算処理にパイプライン処理を適用し、演算間並列処理を実現することで、その性能向上を図っている。しかし、多重結合演算ではパイプライン段数の増加に従い多量のデータがネットワークを流れることになり [1]、分散メモリ型並列計算機環境においてはネットワーク上でのデータ量を考慮した実装方式が必要となる。

本報告では並列多重結合演算を分散メモリ型並列計算機 (Fujitsu AP1000DDV) 上に実装し、Left-Deep 木および Right-Deep 木 [2] を用いた処理方式の評価を行う。

2 AP1000DDV の構成

富士通社製分散メモリ型並列計算機 AP1000[3] は、図 1に示す構成をとる。

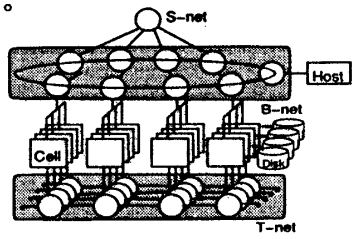


図 1: AP1000 の構成

今回の性能測定では 64 台のセル (SPARC(25MHz),DRAM (16MB)) 構成のシステムを利用した。各セルには、DDV(分散ディスクビデオ) ボードで 1GB のローカルディスクが 1 台接続されている。セル間の 1 対 1 通信は二次元のトーラストポロジを持つ通信ネットワーク T-net、1 対多通信はブロードキャストネットワーク B-net、バリア同期はツリー構造を持つ S-net を利用した。

3 多重結合演算の実装方式

多重結合演算の処理は、木構造を用いて結合演算を行う順序が表現され、木の傾きにより分類される。

今回は AP1000DDV システム上に図 2に示す Right-Deep 木と Left-Deep 木の多重結合演算を実装した。また、結合演算

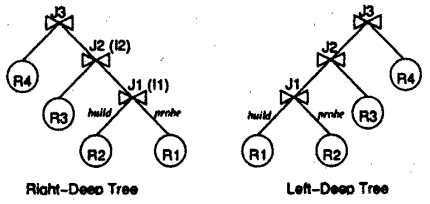


図 2: 多重結合演算のスケジューリング木

の処理方式としてハッシュ結合方式を用いた。ハッシュ結合方式は、ノードの左側のリレーション (ビルドリレーション) にハッシュ関数を適用し主記憶上にハッシュ表を生成するビルドフェーズ、ノードの右側のリレーション (プローブリレーション) にハッシュ関数を適用しハッシュ表上での突き合わせ処理と結合処理を行うプローブフェーズからなる。

3.1 多重結合演算木の処理

本節では、より複雑な Right-Deep 木 (図 2) の処理を示す。各セルに割り当てられたビルドリレーション R_2, R_3, R_4 のハッシュ表は主記憶上に格納できるものとする。

1. 各セルは R_2 をディスクから読みだし、ハッシュ関数を適用し、ハッシュ値により対応するセルに T-net を介して送信する。各セルは受信したタプルのハッシュ表を主記憶上に作成する。同様に R_3, R_4 のハッシュ表を順次作成する。全ての処理が終了した時点で他の全てのセルに B-net を介して終了メッセージを放送する。
2. 各セルはプローブリレーション R_1 をディスクから読み出し、各タプルにハッシュ関数を適用し、ハッシュ値により対応するセルに T-net を介して送信する。各セルは受信したタプルを対応するハッシュ表との突き合わせ処理、結合処理を行う。演算 J_1, J_2 の結果の場合、再びハッシュ関数を適用し、対応するセルに T-net を介して送信する。最終結果である場合にはディスクに書き込む。全てのタプルの処理が終了した時点で S-net を用いて同期をとる。

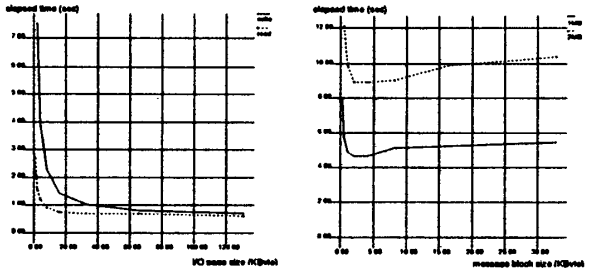
3.2 実装方式

分散メモリ型並列計算機ではディスク入出力処理と通信処理をオーバラップさせ、効率良く処理する必要がある。今回の実装では、ディスクからリレーションを読み出し対応するセルに送信する処理、受信したタプルのハッシュ表の作成 (ビルドフェーズ)、結合演算の結果をディスクに書き込むまたは対応するセルに送信する (プローブフェーズ) 処理の 2 種類のプロセスとした。また、図 3に示すように基本性能測定として行った入出力のページサイズを変化させた入出力時間の結果 (ファイルサイズ 1MB)、及び、セル間の通信サイズ (ブロックサイズ) を変化させた場合の単一結合演算処理時間の結果 (セル構成 8×8、リレーションサイズ 1MB,2MB) から最適値を選択し、ページサイズを 64KB、ブロックサイズを 4KB とした。

4 性能測定

本節では性能評価について述べる。また、測定環境は全てのリレーション、全ての結合演算の結果のタプル長を等しくし、選択率は 1、ソースリレーションのデータの偏りはなく、ソースリレーション、中間結果はセル間で均等に分割されるものとした。また、セル台数は 4(2×2),16(4×4),32(8×4),64(8×8) で測定した。また、データは全て均等にディスクに格納されているものとし、ハッシュ後の分割についても、セル間で等しい大きさになるように分割されるものとする。

Implementation and Performance evaluation of Multi-Join processing on the AP1000DDV.
T.Shintani, M.Nakano and M.Kitsuregawa
Institute of Industrial Science, University of Tokyo
Roppongi 7-22-1, Minato-ku, Tokyo, 106 Japan



ディスク入出力処理時間　　結合演算処理時間

図 3: 基本性能測定結果

4.1 台数効果

セル台数の増加による処理性能の変化を図 4に示す。ここでは、結合演算数 (リレーション数) を 3(4)、リレーションのファイルサイズを 1MB× セル数、タプル長を 256B とし、1 つのセル当たりの処理量を均一とした。図 4は、1 セル上でリレーションのファイルサイズ 1MB における Right-Deep 木、Left-Deep 木の処理時間を測定し、セル数 4,16,32,64 の結果を 1 セルの処理時間で正規化したものである。

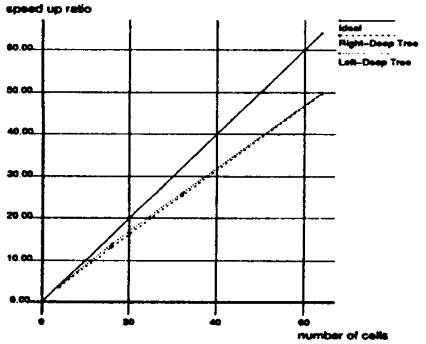


図 4: 台数効果

図 4から Right-Deep 木、Left-Deep 木共に、セル台数が増加するに従い性能向上が見られる。しかし、通信量の増加、セル間の同期オーバヘッドにより、台数の増加に従い性能向上比は低下している。

4.2 タプル長を変化させた場合

タプル長を変化させた場合の処理時間の変化を図 5に示す。入出力サイズは固定としているため、タプル長が短くなると相対的にタプル数が増え、CPU 処理負荷が増大する。ここでは、結合演算数 (リレーション数) を 3(4)、セル台数を 64(8×8)、リレーションのファイルサイズを 64MB とし、タプル長を 64B から 1KB まで変化させた。

図 5では Right-Deep 木の処理時間が Left-Deep 木よりも長くなっている。Right-Deep 木は Left-Deep 木と比べ通信処理の負荷が高い、つまり、複数の結合演算のプローブフェーズパイプライン処理を行っているため、入出力コストに対して CPU 処理コスト、通信コストが大きくなっている。Right-Deep 木、Left-Deep 木共に全体のディスク入出力のコストは等しいため、この差は通信負荷、CPU 処理負荷によるものである。しかし、タプル長が長くなるに従い処理時間の差が減少しており、CPU 処理による負荷の影響が大きいと考えられる。

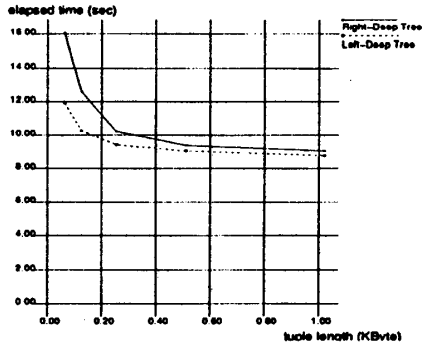


図 5: タプル長を変化させた場合

4.3 結合演算数を変化させた場合

Right-Deep 木、Left-Deep 木の結合演算数を変化させた場合の処理時間の変化を図 6に示す。ここでは、セル台数を 64 (8×8)、リレーションのファイルサイズを 64MB、タプル長を 256B とした。

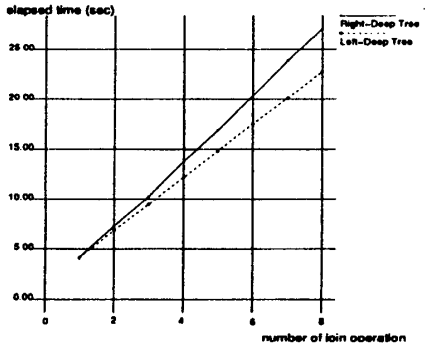


図 6: 結合演算数を変化させた場合

図 6でも Right-Deep 木の処理時間が Left-Deep 木よりも長くなっており、結合演算数の増加に従って処理時間の差が増大している。これは CPU 処理の増加と共に、ネットワーク上でのデータ転送量も増大しているからである。

5 終りに

分散メモリ型並列計算機 AP1000DDV 上に多重結合演算を実装し、性能評価行うことにより、多重結合演算処理の並列化の実装についての検討を行った。並列多重結合演算処理では、システム資源 (ディスク入出力性能、CPU 性能、通信性能) のバランスの良い実行方式を考慮しなければならないことが確認された。

謝辞

この研究を行うにあたり、AP1000 を利用する場を提供していただいた富士通株式会社 大江様、堀江様をはじめとする並列処理研究センターの方々に深く感謝致します。

参考文献

[1] 中野, 新谷, 喜連川: 並列データベースシステムにおける多重結合演算処理の最適化とその評価, 情報処理学会 アーキテクチャ研究会,SWoPP'95 (1995)

[2] Schneider,D.A and DeWitt,D.J: *Tradeoffs in Processing Complex Join Queries via Hashing in Multiprocessor Database Machines*, Proc.of VLDB, pp.469-480(1990)

[3] 清水, 堀江, 石畑: *AP1000* の性能評価 - メッセージハンドリング, 放送, バリア同期 -, SWoPP'92(92-ARC-95-12), 情処研報,Vol.92,No.64 (1992)