スーパーデータベースコンピュータSDCに於ける単一モジュールの基本性能

5H－6

平野 聡　原田 昌信　楊 維康　喜連川 優　高木 幹雄
東京大学 生産技術研究所

1 概要

本論文ではスーパー・データベース・コンピュータSDC [1][2]の単一モジュールの性能評価について報告する。 以下、SDCの特徴について簡単に述べた後、Wisconsin Benchmark[3]により試作版単一モジュールの性能評価を行ない、Teradata社のデータベースマシンDCB/1012[6]及びGAMMA[5]と比較する。

2 SDCの特徴

(1) **ハイブリッド・アーキテクチャ**: 図1にSDCの全体構成を示す。SDCではプロセッサ100台規模の並列性を目指して、プロセッサ4台、磁気ディスク装置2台を密結合の処理モジュールとし、それらを高機能オメガ・ネットワークで疎に結合したハイブリッド・アーキテクチャをとる。この構成では、密結合の利点である軽い通信コストによる高速性と、クラスタ数の増減によるスケーラビリティが同時に得られる。並列性はモジュール内、モジュール間の二つのレベルで実現され、ダイナミックGRACEハッシュアルゴリズム[7]を直接反映している。

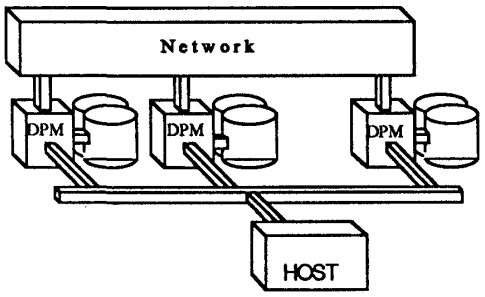

図1: SDCの全体構成

(2) **高速2次記憶系と処理系の結合**: 図2にデータベース処理モジュール内の構成を示す。SDCの大容量2次記憶は高速なディスクを複数台用い、リレーションをタプル単位で水平に分割して格納することにより、ストライピング・ディスクを構成して広いバンド幅を実現する。2次記憶は高速データ転送専用バスH-Busを介してプロセッサPn群と密に結合されており、I/Oとプロセッサ間のボトルネックが解消されている。また、ディスクに近いところで前処理を行なう為、モジュール外への不要なデータの転送を抑えネットワークの負荷を軽減している。

(3) **ハードウエア・ソータ**: 各モジュールは19個のLSIソートチップからなるハードウエアソータを有し、512Kレコードまでのソート処理を線系時間で高速に実行する。大容量ソータメモリはソータとしてもデータメモリDMの拡張としても使用可能であり、ソート機能を有するバイモーダル・メモリとみなすことができる。

(4) **高機能オメガネットワーク**: 高機能オメガネットワークは結合演算の際、データの分布が不均一でもパケットを各モジュールに均等に割り当てるパケット平坦化機能を持つ。ネットワークのスイッチング・ユニットはパケットの分布に関する局所的な情報のみからネットワークの接続状態を決定する。

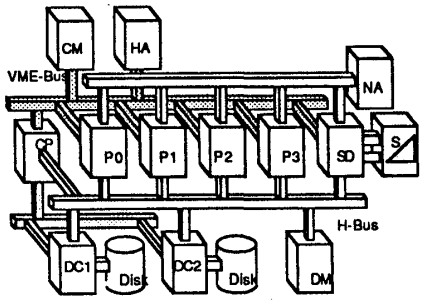

図2: モジュール構成

(5) **ダイナミック・ハイブリッドGRACEハッシュ・アルゴリズムDHGH**: 従来結合演算で用いられてきたハイブリッド・ハッシュ・アルゴリズムでは、不均一なデータ分布を考慮していなかった。DHGHではパケットの動的デステージングにより、パケットのデータ分布に因らず一定の処理時間で結合演算を実行できる。

3 ハードウエア諸元

評価した試作機及び比較対象のハードウエア仕様[5]は以下の通りである。

- Teradata DCB/1012
  - プロセッサi80286 20台
  - 8インチディスク40台
  - 各プロセッサにバッファメモリ2メガバイト、計40メガバイト
- GAMMA
  - プロセッサVAX11/750 17台
  - 8インチディスク8台
  - 各プロセッサにバッファメモリ2メガバイト、計34メガバイト
- SDC (1モジュール)
  - プロセッサMC68020 20Mhz 5台、うち1台はコントロール用
  - 8インチディスク2台
  - バッファメモリ8メガバイト

4 性能評価

評価はウィスコンシンベンチマークを用い3種類のリレーションについて行なった。4バイトのキーを持つ208バイト長のタプルをそれぞれ1万個、10万個、100万個含む。上記の3種のマシンいずれもデータを複数のディスクに水平に分割して格納している。

4.1 選択演算

選択演算では1%または10%の選択を行なった。Teradataでは結果の格納時にリカバリーのための処理を行なっているがGAMMAとSDCでは行なっていない。また今回はインデックスを使用していない。

---

[0]SDC, The Super Database Computer, Basic performance of one module

S.Hirano, W.Yang, M.Harada, M.Kitsuregawa, M.Takagi
Institute of Industrial Science, University of Tokyo

```
insert into result
select * from relA where k1 < X;
```

選択演算	Teradata	GAMMA	SDC
1万タプル 1%	6.86	1.63	0.57
1万タプル 10%	15.97	2.11	0.61
10万タプル 1%	28.22	13.83	4.80
10万タプル 10%	111.0	17.44	5.23
100万タプル 1%	213.1	134.9	47.8
100万タプル 10%	1107	181.7	52.0

(単位 秒)

図3はリレーションサイズの変化に対するSDCの実行時間の変化を表している。SDCはストリーム指向の処理方式をとるので、CPUの処理はディスクの読み出しとオーバーラップしている。入力リレーションだけでなく出力ファイルもディスクストライピングにより2つのディスクに分散して格納し、入出力時間の短縮を図っている。

図4に選択演算でのプロセッサの台数効果を示す。リレーションは100万タプルで、横軸は選択率、縦軸は実行時間を表している。プロセッサ数が2台～4台の場合は全ての選択率にわたってディスクの入力速度に追従しているが、プロセッサ数が1台の時には選択率50%以上になると、処理が追従不能になり急激に性能が低下して行く。

4.2 選択演算

結合演算 (joinAselB) の結果を示す。入力リレーション relB は結合前に10%の選択演算を施される。

```
insert into result
select A.*, B.* from relA A, relB B
where A.k1 = B.k1 and B.k1 < X;
```

結合演算	Teradata	GAMMA	SDC
1万タプル	35.6	5.1	1.14
10万タプル	331.7	36.3	10.5
100万タプル	3535	703.0	127.0

(単位 秒)

今回の評価では演算アルゴリズムとして hyblid hash join[4] を用いた。Teradata では Hash sort-merge join、GAMMA では Simple hash join を用いている。図5はリレーションサイズの変化に対する実行時間の変化を表している。これからタプル数の増加に対して線形の時間で終了していることがわかる。Teradata と GAMMA ではプロセッサ同士がネットワークを介してタプルを交換するのでオーバーヘッドが生ずるのに対して、SDC では密結合の利点を生かしてディスクと各プロセッサがメモリを共有し、前処理の選択演算、ハッシュ、つき合わせ操作を行なうので、ディスクの入出力時間とオーバーラップして内で処理を行なっている。

5 結論

ベンチマークにより SDC ではプロセッサ台数、ディスク台数共に Teradata、GAMMA と比べてわずかであるにもかかわらず、非常に高い性能を持つことが明らかになった。現在はこのモジュールを構成要素とする多モジュール版の実装を進めている。

参考文献

[1] 楊、平野、喜連川、高木「スーパーデータベースコンピュータ SDC のアーキテクチャ」情報処理学会第39回全国大会、1989

[2] 平野、楊、喜連川、高木「スーパーデータベースコンピュータ SDC のシステムソフトウエアの概要」情報処理学会第39回全国大会、1989

[3] Bitton,D.,DeWitt,D.,C.Turbyfill "Benchmarking Database Systems - A Systematic Approach" Very Large Databases Conf., 1983

[4] DeWitt,D.,R.Gerber, "Multiprocessor Hash-Based Join Aligolithms," Very Large Databases Conf., 1983

[5] DeWitt,D.,Ghandeharizadeh,S.,et al, "A SINGLE USER EVALUATION OF THE GAMMA DATABASE MACHINE" 5th International Workshop on Database Machines, 1985

[6] Teradata Corp., "DBC/1012 Data Base Computer Concepts & Facilities" Teradata Corp. 1983

[7] Kitsuregawa,M.,Nakayama,M.,Takagi,M. "The effect of bucket size tuning in the dynamic hyblid GRACE hash join method" Very Large Databases Conf., 1989

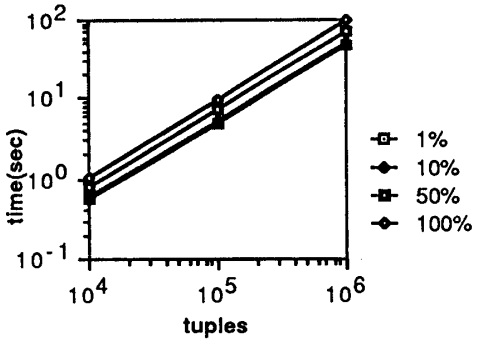

図 3: 選択演算

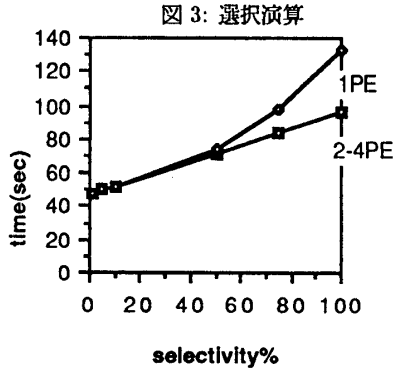

図 4: 選択演算の台数効果

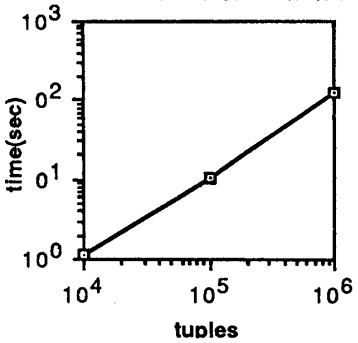

図 5: 結合演算