# 3B－1　スーパーデータベースコンピュータ SDC-II における不均一データ分布での結合演算処理に対する性能評価

中村 稔　田村 孝之　喜連川 優　高木 幹雄
東京大学　生産技術研究所

## 1　はじめに

スーパーデータベースコンピュータ SDC-II は関係データベースシステムにおける問い合わせ処理の超高速実行を目的としたバックエンド型高並列 SQL サーバである．SDC-II は密結合型データ処理モジュール並びに複数の処理モジュールを疎に結合する高機能オメガネットワークからなるハイブリッドアーキテクチャを採る．

SDC-II のモジュール間を接続するデータネットワークはオメガトポロジーを採用し，通常のデータ送付先モジュールを指定した宛先指定モードの他にバケット平坦化モードを支援する．この機能により特定の処理モジュールによる集中制御なしに複数のモジュール間の負荷の均等化を行なうことができる [2]．

本論文では，SDC-II 上での不均一データ分布下における結合演算の性能評価結果について報告する．

## 2　SDC-II における結合演算処理

SDC-II では単一結合演算として GRACE ハッシュ結合演算と平坦化ハッシュ結合演算を実装した．GRACE ハッシュ結合演算アルゴリズム [1] はハッシュ分割後のバケットの大きさが不均一であった場合，これが結合演算フェーズにおけるモジュール間での負荷の偏りとなり，処理速度の低下をもたらす．

平坦化ハッシュ結合演算アルゴリズム [2] ではモジュール間の均等な負荷分散を実現するためにバケット分割の際にバケット平坦化を行なう．このため不均一なデータ分布に対しても良好な性能を得ることができる．

## 3　性能評価

SDC-II 上に実装された GRACE ハッシュ結合演算および平坦化ハッシュ結合演算に関して性能評価を行なった．性能評価は表 1 に示す環境下で行った．

性能評価は拡張ウィスコンシンベンチマークに準じ，タプル長 208 Bytes，タプル数 100 万件 × モジュール数のリレーションに対して結合演算処理の実行に要する処理時間を測定した．入力リレーションには結合前に 10%の選択演算が施される．

| | |
|---|---|
| モジュール数 | 1 ～ 6 |
| DP 数/モジュール | 4 |
| データネットワーク最大転送速度 | 10MB/sec |
| ディスク数/モジュール | 4 |
| ディスク容量/ディスク | 520MB |
| 平均シーク速度 | 12ms |
| 最大転送レート/ディスク | 3.05MB/sec |

表 1: 性能評価環境

モジュール間の負荷の偏りをシミュレートするために，それぞれのアルゴリズムでバケット分割時に各バケットに含まれるタプルの数が Zipf 分布に準ずるように振り分ける．

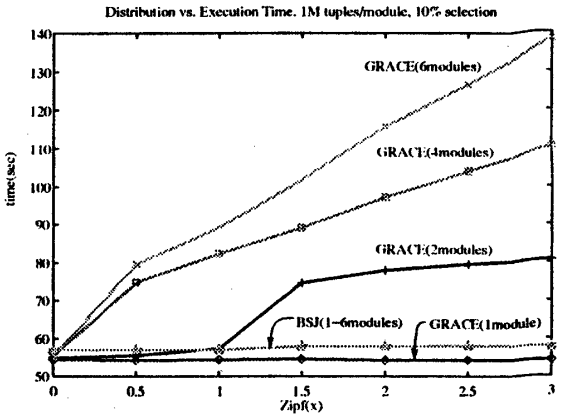


図 1: 不均一分布データに対する処理性能

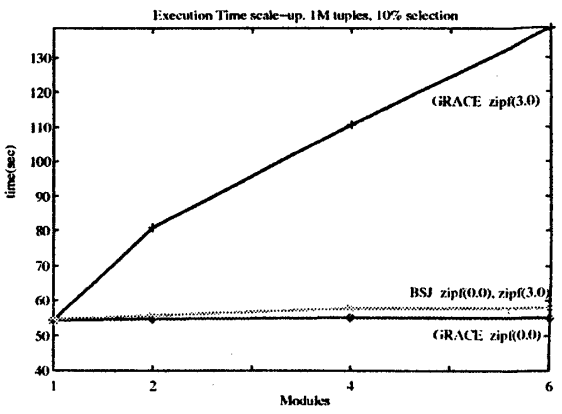


図 2: 結合演算実行時間

Performance evaluation of Join processing with data skew on the Super Database Computer:SDC-II
M.Nakamura, T.Tamura, M.Kitsuregawa and M.Takagi
Institute of Industrial Science, University of Tokyo

図 1に SDC-II においてサブバケットのタプル数を不均一分布にした場合のベンチマークの実行結果を示す.

図 1から平坦化ハッシュ結合演算が負荷の偏りに関わらずほぼ一定の時間で処理を完了していることがわかる. これは平坦化ハッシュ結合演算がバケットの平坦化によるモジュール間の負荷の均一化をおこなうためで, このような処理を行なわない GRACE ハッシュアルゴリズムではデータの分布の偏りが大きくなるにしたがって処理が遅くなる.

SDC-II では各モジュールが 32MByte のステージングバッファを持つので, データの分布が均一であった場合はメモリ上で結合演算処理が実行できる (Zipf(0) の場合). データの分布が不均一の場合 GRACE ハッシュアルゴリズムでは特定のモジュールにタプルが集中する. これを中間ファイルとしてディスクに書き出すため 2 モジュールの場合 Zipf(1.5) 以降で, また 4 および 6 モジュールの場合 Zipf(0.5) 以降で処理性能が大きく低下している.

図 2にモジュール数を変化させた時の実行時間の変化を示す.

GRACE ハッシュアルゴリズムでは負荷に偏りがあるとモジュール数の増加するにしたがって負荷が特定のモジュールに集中するために全体の処理速度の低下につながるのに対して, 平坦化ハッシュアルゴリズムでは常に一定の処理性能を示していることがわかる.

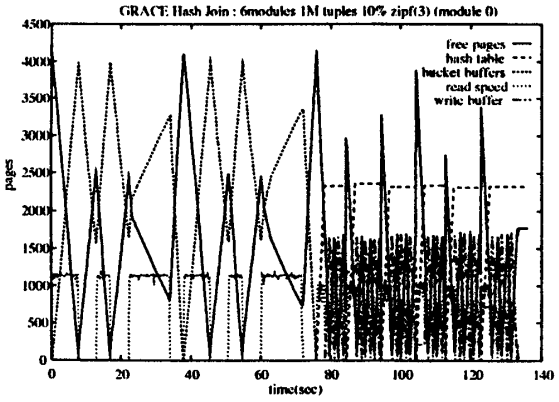


図 3: GRACE ハッシュ結合演算 (DPM 0)

図 3に GRACE ハッシュ結合演算を Zipf(3) のデータ分布の元で実行した場合のモジュール 0 のメモリの使用状況を, 図 4に同じ処理を実行時のモジュール 5 のメモリの使用状況を示す.

データ分布が Zipf(3) の場合全タプルのうち約 84% がモジュール 0 に収集され, 逆にモジュール 5 には全体の約 1.5%だけが収集される. モジュール 5 ではデータのほとんどをモジュール 0 に送り出すためフリーページがあまり消費されない. システム全体のメモリの使用効率が悪いために全体の処理性能が低下しているといえる.

図 5に平坦化ハッシュ結合演算を Zipf(3) のデータ分布の元で実行した場合のモジュール 0 のメモリの使用状況を

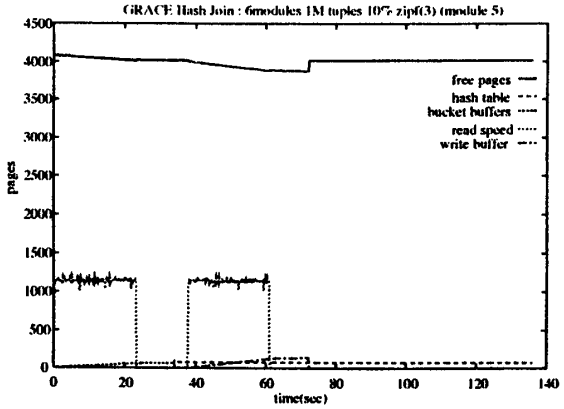


図 4: GRACE ハッシュ結合演算 (DPM 5)

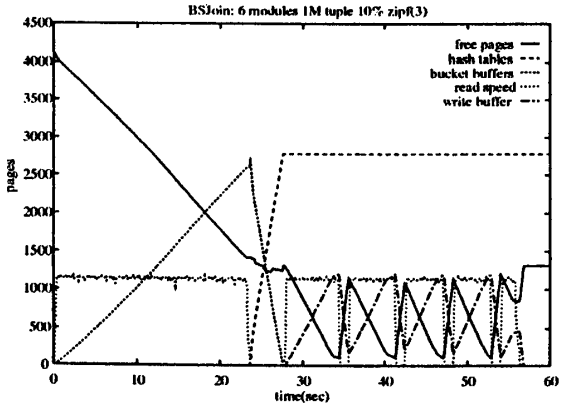


図 5: 平坦化ハッシュ結合演算 (DPM 0)

示す. 平坦化ハッシュ結合演算ではデータ分布が負均一であっても全てのモジュールで同様の振舞を示す. これは平坦化ハッシュ結合演算が均一に負荷をモジュール間で分散していることを表している.

## 4 まとめ

本論文では SDC-II における結合演算処理に関し, 不均一データ分布下での処理性能を示した. データ分布が不均一な環境下においては平坦化ハッシュ結合演算による負荷分散の有効性が確認できた. 今後更に詳細な性能評価を行なう予定である.

## 参考文献

[1] Kitsuregawa, Tanaka, and Moto-oka. Application of Hash to Data Base Machine and Its Architecture. *New Generation Computing*, 1983.

[2] 中村, 平野, 原田, 相場, 鈴木, 喜連川, 高木. スーパーデータベースコンピュータ (SDC) 上での平坦化ハッシュジョインの評価. 並列処理シンポジウム *JSPP'92*, 1992.