7W－2

スーパーデータベースコンピュータ SDC-II における ライトディープ多重結合演算の評価

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　概要

SDC-II は非定型問合せを高速に実行する高並列のバックエンド型 SQL サーバである．SDC-II は従来開発してきた SDC-I の経験に基づきその拡張を試みた試作機であり，処理モジュールの性能向上と高機能データネットワークによる複数モジュールの相互接続および多モジュール構成下での評価を行なっている．本論文では SDC-II 上におけるライトディープ多重結合演算処理の実装とその評価結果について述べる．

2　SDC-II の構成

図 1 に SDC-II の構成を示す．

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．このような構成により密結合の利点である軽い通信コストによる高速性とモジュール数の増減によるスケーラビリティが同時に得られる．

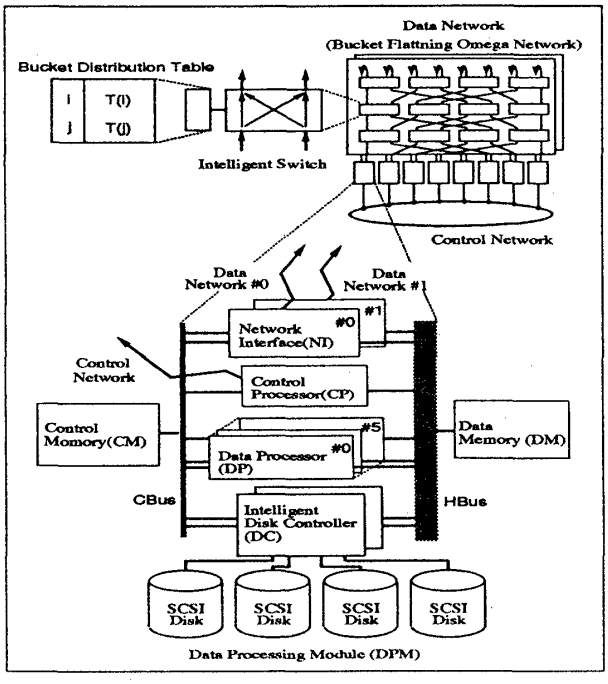

Figure 1: SDC-II の構成

Evaluation of Right-Deep Multi-way Join on the Super Database Computer:SDC-II
M.Nakamura, T.Tamura, M.Kitsuregawa and M.Takagi
Institute of Industrial Science, University of Tokyo

3　多重結合演算の実装

複数の結合演算からなる多重結合演算はハッシュテーブルを生成するビルドフェーズとハッシュテーブルを走査し結果を生成するプローブフェーズの組合せによって Right-Deep, Left-Deep およびそれ以外の Bushy Tree に分類される．

今回は SDC-II 上に Right-Deep 多重結合演算を実装した [2]．

Right-Deep 多重結合演算では複数のハッシュテーブルに対するプローブ処理を並行して実行するため結合演算処理の負荷が高く，またシステムの制御も複雑なものとなる．SDC-II では Right-Deep 多重結合演算を実装するため複数のハッシュテーブルの動的な割り当てと解放，複数のネットワークバッファと複数のモジュールに跨るデータ流制御とデッドロック回避機構，複数の netInBuffer からのページの取得と netOutBuffer への EOF の伝達などの機能拡張を施した．

4　性能評価

SDC-II のシステムソフトウェア上に Right-Deep 多重結合演算を実装し性能評価を行なった．

性能評価には ドラフト版の TPC-D ベンチマーク [1] から各々 2 多重，3 多重および 4 多重の多重結合演算を含む Query3, Query5 および Query9 を用いた．ただし，Query3 および Query9 に関しては処理の都合上修正を加えている．例として図 2 に Query 5 の内容を示す．

```
select
  C_Nation,
  sum(L_Extendedprice*(1-L_Discount/100))
from
  customers, orders, lineitem, suppliers
where C_Custkey = O_Custkey
  and O_Orderkey = L_Orderkey
  and L_Suppkey = S_Suppkey
  and C_Nation = S_Nation
  and S_Region = 'ASIA'
  and O_Orderdate >= date("1/1/1994")
  and O_Orderdate < date("1/1/1995")
group by C_Nation
order by 2 desc
```

Figure 2: TPC-D ベンチマークにおける Query5

使用したデータベースは表 1 に示すような TPC-D 用のものをスケールファクター 1 (データベースサイズ:約 100MB) から 4 (データベースサイズ:約 400MB) の範囲で使用した．実行環境は表 2 に示す通りである．

図 3 は単一モジュールにおけるスケールファクターに対する処理速度の変化を示したものである．各 Query とも処理時間がほぼスケールファクターに比例していることがわかる．

リレーション	タプル長	タプル数
customers	180bytes	15000 × *ScaleFactor*
orders	112bytes	150000 × *ScaleFactor*
lineitem	156bytes	600000 × *ScaleFactor*
parts	164bytes	20000 × *ScaleFactor*
partsupp	144bytes	80000 × *ScaleFactor*
suppliers	160bytes	1000 × *ScaleFactor*

Table 1: Test relation for TPC-D

モジュール数	1 ～ 4
DP 数/モジュール	3
ディスク数/モジュール	4
最大読み出し速度/ディスク	2.47Mbyte/sec
データネットワーク数/モジュール	1
最大転送速度/データネットワーク	10MB/sec
ステージンバッファ/モジュール	32MB

Table 2: 性能評価環境

SDC-II ではモジュール内のプロセス間におけるデータ交換はバッファと呼ばれるデータ構造を通じて固定長のページを受け渡すことで実現している. 図 4 は Query 5 を 4 モジュールで実行した際の DPM 0 上のバッファの使用状況とディスクからの読み出し速度を記録したものである. 横軸は経過時間, 縦軸はバッファ中のページ数を表す. 但し "read speed" に関しては 1 秒あたりにディスクから読み出すページ数の合計である. また 横軸に沿って各リレーションの処理期間を示している.

図 4 において "orders" のビルドフェーズでは read speed がディスクの読み出し速度に近い値を示しており, DP の処理がディスクからのデータに追い付いていることがわかる. また, プローブフェーズの様に負荷の高い処理では readBuffer が上限値に達し, ディスクからのデータの流入を押えている様子がわかる.

図 5 はモジュール数に対する処理速度の変化を Query3, Query5, Query9 の各々について示したものである. モジュール数に応じてスケールファークターを増やしているため理想的な状態ではグラフは水平になる. 複数モジュールでの処理においてはモジュール間でタプルの交換が行なわれるため, 処理すべきデータ量が同じであっても負荷はモジュール数の増加に従って大きくなる. Query5, Quer9 については複数モジュールになると DP での処理が追い付かなくなるためモジュール数 2 を境に処理速度が低下している. Query3 は他に比較して負荷の軽い処理であるため複数モジュールでの処理においても過負荷による性能低下はない. Query3 における性能低下は複数モジュールにおける同期のオーバーヘッドによるものである.

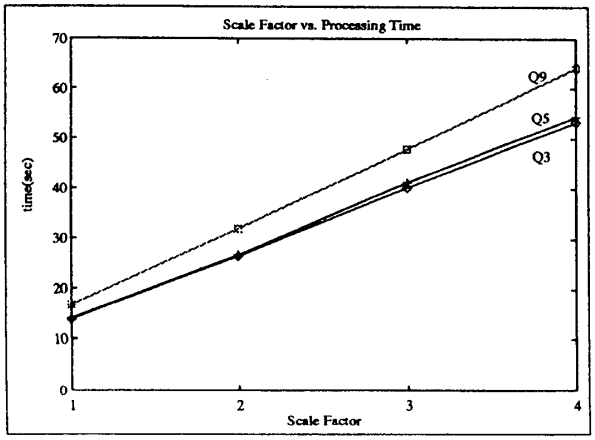

Figure 3: Scale Factor vs. Processing Time

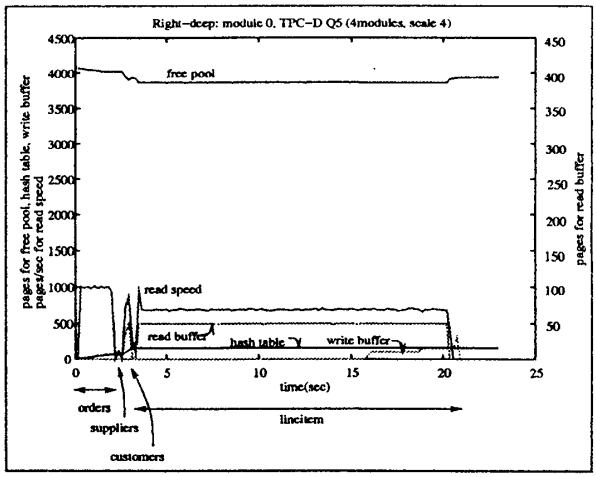

Figure 4: Buffer trace of Q5: 4 modules, Scale Factor 4

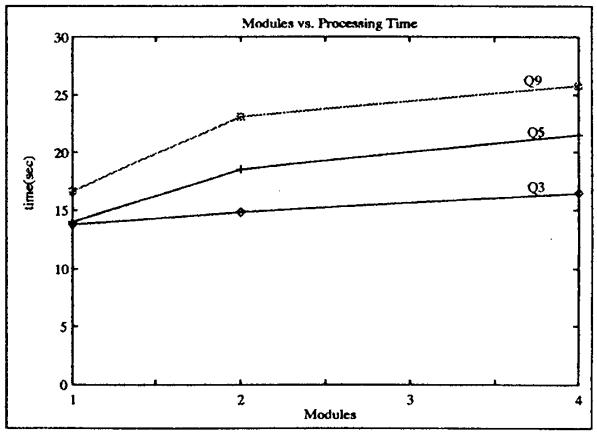

Figure 5: Modules vs. Processing Time

5 まとめ

本論文では, SDC-II 上における多重結合演算の試作機上における実行結果を示した. TPC-D ベンチマークによる評価結果から SDC-II のハードウェアならびにシステムソフトウェアが究めて効率良く稼働していることがわかった. 今後更に詳細な性能評価を行なう予定である.

References

[1] *TPCBenchmark^{TM} D(DecisionSupport)* Working Draft 6.0. Transaction Processing Performatce Council. 1993.

[2] 中村稔, 田村孝之, 喜連川優, 高木幹雄. スーパーデータベースコンピュータ SDC2 における多重結合演算の実装と評価. *SWoPP '94*