GRACEに於けるモジュール間結合方式

4P-2

坂井　修一　　喜連川　優　　田中　英彦　　元岡　達

（東京大学　　工学部）

1．はじめに

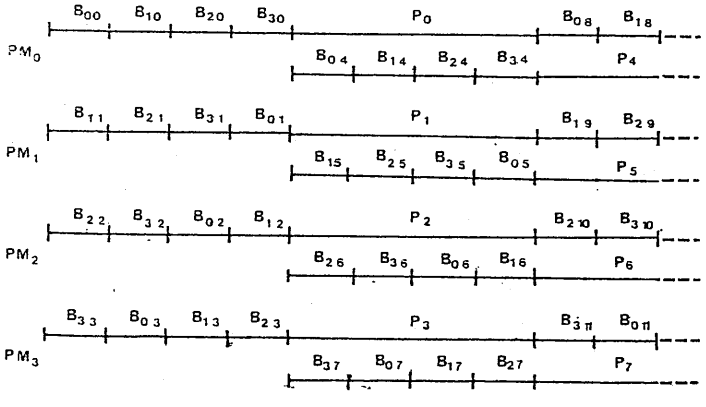
GRACEでは、プロセシングモジュール（PM）、メモリモジュール（MM）、ディスクモジュール（DM）が各々複数台結合したアーキテクチャ上で高速な関係代数演算を実現している。これらモジュール間の結合方式として、従来から光ファイバを用いた時分割多重チャネル方式のリングバスが考えられてきたが，今回は多段スイッチネットワークでの実現方式について報告する。

多段スイッチネットワークの種類も、クロスバ・スイッチのようにハードウェア量の大きなものから、オメガネットのようなものまで様々であるが、GRACEの処理方式の特徴を活かした、$\log_2 N$段のネットワークについて検討した。

2．GRACE上でのパイプライン処理

実際に関係代数演算を実行する手順は次の通りである。（ただし、PM内のソータは、途切れることなく次々とオペレーションを処理するパイプライン・マージソータを用いるとする。）

1）．当該リレーションをDMからMMにステージする。その際、ハッシュをかけ、ハッシュバケットはMM側で均等な大きさになるように分割する。



P_i : processing of i-th bucket
B_{ij}: input of j-th bucket in MM_i

Fig.1. pipeline processing overview

2）．必要に応じてバケットサイズのチューニングをする。バケットは小さい順にソートする。

3）．1つのPMが1つのハッシュバケットを取り込んで処理する。結果リレーションは、次の演算に使用する場合には再びハッシュを施す。この時のMMとPMの結合の様子をFig. 1に示した。

4）．結果リレーションをMMに戻す。さらに必要な時にはDMに戻す。

3．MM to PM ネットワーク

Fig. 1に見られるように、MMとPMの結合は極めて規則的である。即ち、1つのオペレーションのみに注目すれば、パーミュテーションのクラスは、サーキュラ・シフトが基本となる。サーキュラ・シフトの実現は、$\log_2 N$段ネットワークで可能である。

Theorem. 1.

indirect binary n-cube ネットワーク（Fig. 2.）上では、閉塞なしにサーキュラ・シフトが実現可能である。

（ 証 明 略 ）

従って、indirect binary n-cube ネットワークを使うことを考えた。1回の結合で数KB以上のデータを転送することから、回線交換方式を採用することにした。

実際には、バケットサイズのチューニングには限度があり、閉塞が起こる場合が生ずる。シミュレー

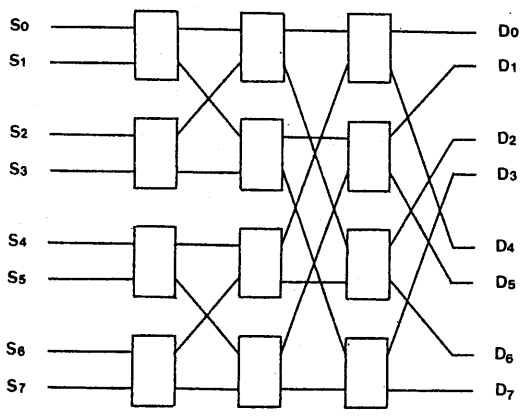


Fig. 2. indirect binary 3-cube network

ションによって調べたところ（Fig. 3.）、閉塞による性能低下は平均0．2%以下、バケットサイズの乱れによる性能低下は平均2%以下に押えられることがわかった。尚、2－2）で述べたバケットサイズ・ソートの効果は非常に大きく、場合によっては20%以上の効率向上をもたらすことが見てとれる。但し、各バケットはMM間に完全に均等に分配されているとした。

複数のオペレーションが同時進行するケースでは、モジュールの割り付けを最適化する必要がある。オペレーション毎に、ネットワークの論理的なパーティショニングを実現し、オペレーションどうしの相互干渉が起きないような環境にする。

Theorem. 2.

N×N indirect binary n－cube [n = $\log_2 N$] ネットワーク上で、$N/2^k$（$0 \leqq k \leqq n$）コのソース｛S^j｝と、$N/2^k$コのデスティネーション｛D^j｝が使用可能のとき、これらが論理的にパーティショニングされたネットワーク上にあるための必要十分条件は、

$\forall j_l \quad S_{i_l}^{(j_l)} = \text{const}. , \quad D_{i_l}^{(j_l)} = \text{const}.$

$(l = 1, 2, 3, \cdots\cdots, k)$

かつ、

フィックスするスイッチが他のオペレーションによって閉塞されていないこと

である。

但し、$S_{i_l} = S_{i_l}^{(n-1)} S_{i_l}^{(n-2)} \cdots\cdots S_{i_l}^{(0)}$（二進表示）とした。

（証明略）

パーティショニングは、大きな処理負荷のオペレーションの次に小さな処理負荷のオペレーションが来たときには容易だが、逆の場合は不可能なケースがある。このときには、パーティションできるまで次のオペレーションを待たせるか、閉塞を承知で実行させるかをコントローラが判断せねばならない。

4．PM to MM, DM to MM ネットワーク

2－1）より、モジュール間の結合は、各タプル毎に変わり、3．のような規則性が無い。また、スループットが主な問題となる環境である。この2点から、データをパケット化し、各スイッチングノードにバッファを設けた$\log_2 N$段ネットワークでの実現が考えられる。

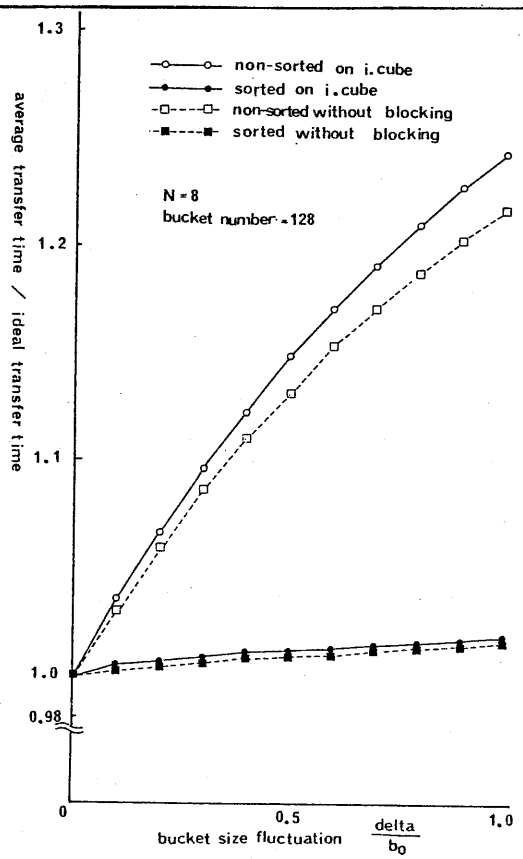


Fig.3. Data transfer time of 1 operation(simulation)

また、バケットの均等な分割を実現するためには、各タプル毎に行先番地を決定する機構が必要である。

複数オペレーションの環境下におけるネットワークのパーティショニングは、MM to PMほど問題にはならない。

尚、コントローラによるモジュールの割り付けは、3，4の2つのネットワークの状態を考慮せねばならない。

5．おわりに

多段ネットワークを用いたGRACEの処理方式の実現について述べた。今後は、ハードウェア・レベルでの詳細な検討を行い、リングバス方式との比較をするつもりである。

参考文献

[1]．喜連川他，「Hash と Sort による関係代数マシン」，信学技報，1981

[2]．Tse-Yun Feng，A Survey of Interconnection Networks，Computer，Dec. 81