4F-2

GRACEに於けるモジュール間結合網とその評価

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

データベースマシンGRACEの構成要素であるディスクモジュール（DM）、メモリモジュール（MM）、プロセシングモジュール（PM）を結合する相互結合網として、時分割多重チャネル方式の光ファイバリングバスや多段スイッチ網が考えられる。

前回［1］はMMからPMにデータを転送する網の方式と性能評価を中心に報告したが、今回はDM、PMからMMにデータをステージングする網に関する制御方式と性能評価を中心に述べる。

2．GRACE上の相互結合網

GRACE上の結合網は以下の2種に大別される。

1）．バケット収集網

MM to PM　網。PMが一つのハッシュバケットを複数のMMから取り込む。両者の結合はサーキュラ・シフトを基本とする規則的なものである。

2）．バケット分配網

PM to MM、DM to MM　網。当該リレーションをMM上にステージする。ハッシュバケットをMM間で均等に分割するため、各タプルの行先を制御する機構が必要である。

3．バケット集収網

結合の規則性から、間接キューブ網を回線交換で用いることを検討した。実際には、バケットサイズの乱れと、MM間でのバケットサイズのゆらぎによって、転送のオーバヘッドが生じる。（前者に関しては既に［1］で述べた。）

後者のもたらす性能低下に関して、シミュレーションによって調べたところ（図1）、MM間でのバケットサイズのゆらぎが大きい時には、20%から30%の転送オーバヘッドが生じることがわかった。また、これが小さい時には、間接キューブ網と完全結合網の差は殆ど無いが、大きくなるにつれて転送時間に開きができる。

4．バケット分配網

バケット分配網では、各バケットをMM間で均等な大きさになるように分配することが重要である。収集網の場合と異なり、結合に規則性が無いので、データ転送のオーバヘッドもバケット収集網に較べて大きいと考えられる。

$\log_2 N$段の網である間接キューブ網を回線交換及び蓄積交換方式で用いる場合について比較検討した。行先アドレスを一様乱数で与えるシミュレーションによって調べた結果を図2に示す。回線交換方式では、網を3枚並列に用いることで、完全結合とほぼ同じ性能を出せることがわかった。その時の転送のオーバヘッドは、64×64の網の場合で80%弱である。蓄積交換方式（網は1枚）では、各スイッチング・ユニットに置かれたバッファの大小が、転送時間に著しい影響を及ぼす。64×64の網の場合、バッ

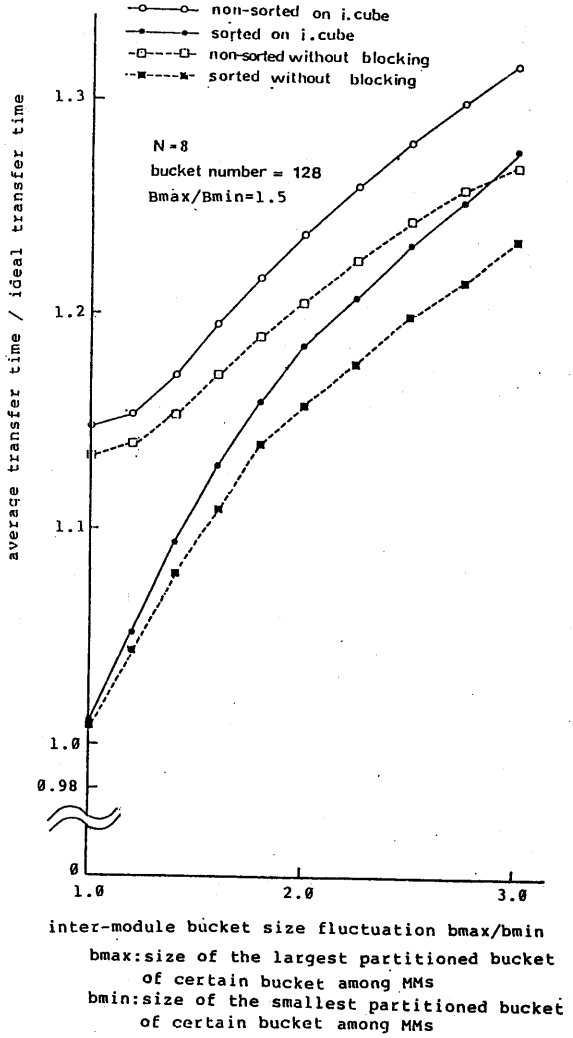

Fig.1 Data transfer time of 1 operation v.s. inter-module bucket size fluctuation

ファサイズ４では転送のオーバヘッドは８０%弱、バッファサイズ１０では５０%程度である。

タプルの行先制御と転送は重畳化して行う。従って、転送時間が長く制御に時間がかけられる時には、１回に転送する全タプルの行先をすべて異ならしめる制御を行って、オーバヘッドを減らすことが考えられる。

この制御を施した場合について、先と同様のシミュレーションを行った（図３）。図からわかるように、回線交換方式では大きな転送時間の短縮が見られる（64×64網の場合、３枚用いればオーバヘッドは約１０%）が、蓄積交換方式で効果的なのは網のサイズが小さい時のみである。

尚、この場合にはMM間でのバケットサイズのゆらぎが生じるが、シミュレーションによって、ゆらぎはたかだか数%にとどまることが判っている。

以上のことから、制御オーバヘッドが支配的でない場合は、先の制御を行った上で回線交換間接キューブ網を３枚程度用いるのが良く、そうでない場合には蓄積交換間接キューブ網を用いるのが有利であると結論される。

他にベネス網、バイトニックソート網、多重チャネルリングバス等による実現が考えられる。これらは結合の実現性という点ではほぼクロスバスイッチと等価であるが、ハードウェア量と制御の手間に差がみられる。

5．おわりに

GRACE上の相互結合網の実現方式とその評価に関して述べた。今後は実際の選択に必要な、コスト面も含めた詳細な検討を行う積りである。

<<参考文献>>

［１］坂井他、「GRACEに於けるモジュール間結合方式」第２５回情処全国大会　１９８２

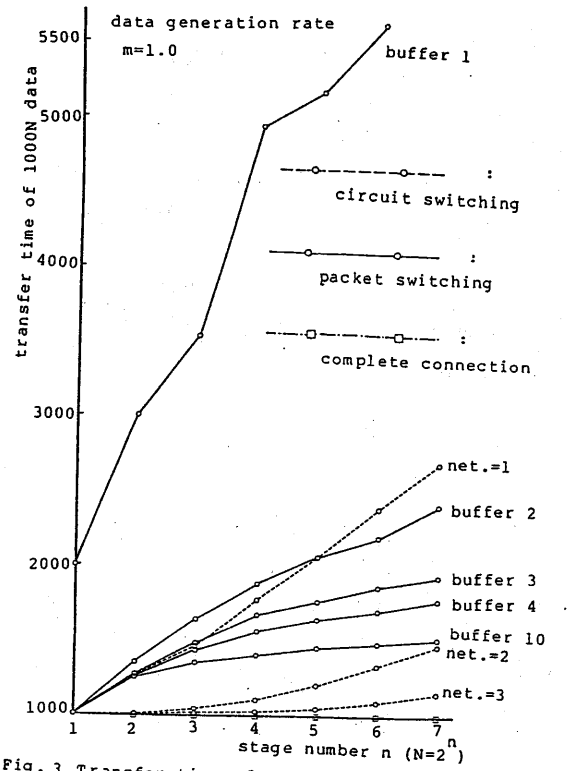

Fig. 3 Transfer time of 1000N data v.s. network size
mapping modified(1:1)
(indirect binary n-cube)

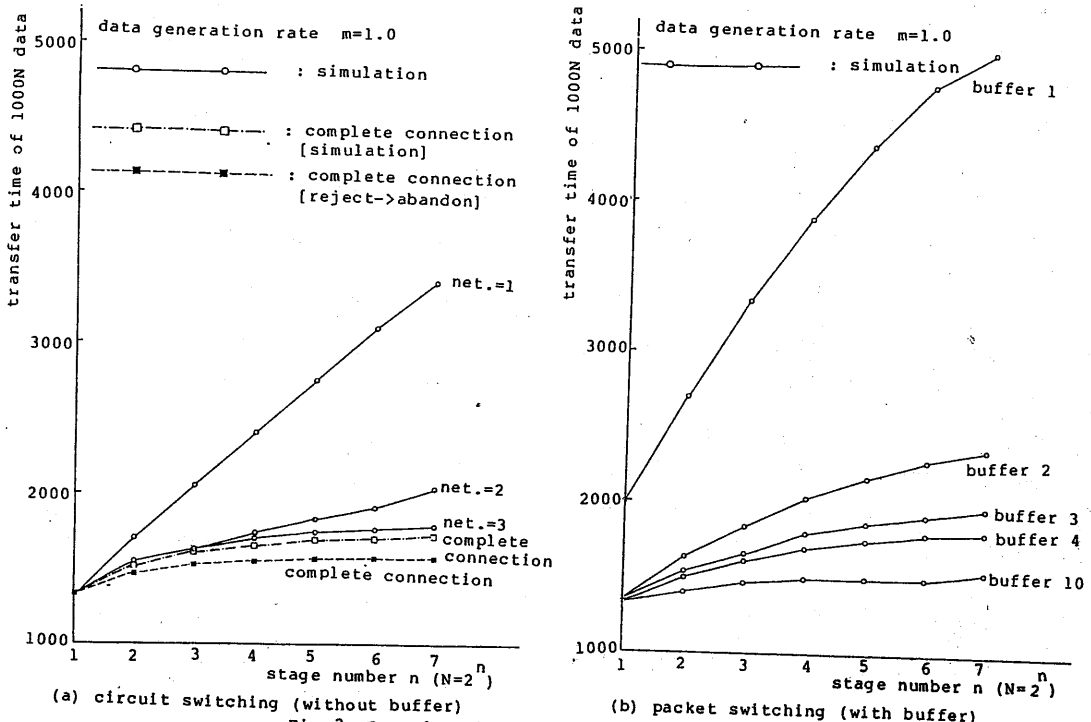

Fig.2 Transfer time of 1000N data v.s. network size
(indirect binary n-cube)