1D-7

# データベースマシンGRACEの モジュール群制御方式

中山　雅哉[†]　伏見　信也[†]　喜連川　優[‡]　田中　英彦[†]　元岡　達[†]

([†]東京大学　工学部　[‡]東京大学　生産技術研究所)

## 1．はじめに

関係代数演算の高並列処理を目的とするデータベースマシンGRACEは，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），コントロールモジュール（CM）の各モジュール群と，これらを結合するリングバス結合網により構成されており，現在その実装を進めている。GRACEには，2種類のデータ転送形態があり，一方の伝送制御手順については，既に検討・評価を行ってきた〔1〕。本稿では，これに引続き，他方の伝送制御方式の検討を行ったので，報告する。

## 2．GRACEに於けるデータ転送形態

GRACE上では，2種類のデータ流が存在し，その転送パタンは互いに異なっている。

（a）PM→MM（バケット分配）　当該処理の結果タプルに対して，次処理の属性に関するhashを施し，割付られたMM群に向かってタプルを送出する。同一のhash値を有するタプルの集合をバケットと呼び，各バケットを構成するタプルは，各MMにできる限り均等に分散される。1つのバケットを各MMに分散させたタプルの集合を，サブバケットと呼ぶ。

（b）MM→PM（バケット収集）　PMが演算処理を実行する為には，各MMのサブバケットを収集して，1つのバケットに戻す必要がある。これは，MM群が各PMに対して巡回的に，サブバケットを転送することによって実現することができる(図1)。

（a）の転送方式に関しては，［1］等で検討・評価されており，以下では　（b）について述べる。

## 3．バケット収集系に於けるスケジュール

バケット収集系においては，上記のように，各MMに分散しているサブバケットを，1つのPMが収集し，バケット単位で処理を行う方法をとっている。ここで，PM（特にSorter）の使用効率を向上させる為には，1つのバケットの大きさを，PMの容量以下で，かつ，それに近くすればよい。

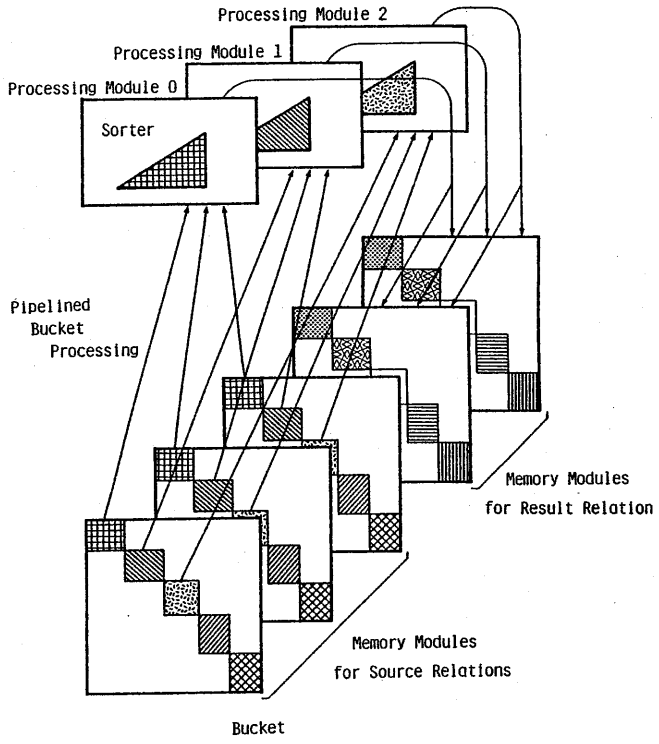


Figure　1　Conceptual View of Pipelined Bucket Processing

しかし，この操作をバケット分配時のhash操作だけで実現することは，非常に困難であるので，バケット分配系におけるhash関数は，PMの容量以下のバケットを数多く生成するように選び，バケット収集の際に，複数のバケットを統合して（プロセシングクラスタと呼ぶ），PMの容量に近くして転送する方法をとることにした。

また，GRACEの処理対象は，バケットシリアルなデータ流であり，この流れの速度によりシステムの性能が決定される。MMからの流れをできる限り乱さずに処理するためには，プロセッシングクラスタの大きさの昇順に転送する必要がある［2］。このように，バケットを統合してプロセッシングクラスタ群を生成し，さらに，これらの転送順序を決定することを，バケット収集系におけるスケジュールと呼ぶ。

データの転送は，スケジュールが決定した時点で，開始することができる。転送の方法として，MM群が，PM群

への転送順序を，巡回的に決めておけば，転送先PMへの転送可能性を，各MMが判断するだけで，並列に，しかもパイプライン的に，行うことができる。

## 4．バケット収集系でのスケジュール制御方式

一般的に，スケジュール制御は，スケジュールの為の正確な情報を，特定のモジュールが，集中管理することによって，実現される。バケット収集系においては，全体のバケットの分布を，転送制御モジュール（Transfer Controle Module）を用いて，集中管理する方法が考えられる（方法１）。

このように，スケジュール専用のモジュールを設けることは，資源管理・通信制御等のオーバヘッドを生ずることになり，不利な点も多い。そこで，各MMがバケット分配時に，自分のサブバケットサイズの分布状況を保持する Hash Count Table の他に，全体のバケットサイズの分布状況を保持する HCTtotal を用いるようにすれば，TCMを用いず，しかも，集中管理を行うことなくスケジュールを実現することもできる（方法２）。

また，バケット分配系で用いられる HCTは，各MMに分散されるサブバケットの大きさを均等に保つのに使われる為，そのサブバケットの間の比率は，HCTtotal におけるバケットの間の比率にほぼ等しくなる。この特徴を用いれば，HCTtotal を用いないで，スケジュールを行うことも可能となる（方法３）。

上記３つの方法による伝送手順を以下に示す。

方法１：TCMを用いたスケジュール制御

（ｉ）　各MMの HCTを集計し，プロセシングクラスタへの統合及び，その転送順序のスケジュールを行う。

（ii）　各MMに，スケジュール表を送信し，転送の開始を指令する。

（iii）　各MMは，スケジュール表とMM内の相対位置をもとに，目的のPMを求め，転送が可能であることを確認して，サブバケットの転送を行う。

方法２：HCTtotal によるスケジュール制御

（ｉ）　各MMは，バケット分配系の転送時に，全体のバケットの分布状況 HCTtotal を個々に保持するようにする。

（ii）　バケット収集の転送命令がCMから送られると，各MMは，プロセシングクラスタへの統合及び，その転送順序のスケジュールを行う。

（iii）　各MMは，スケジュール表とMM内の相対位置をもとに，目的のPMを求め，転送が可能であることを確認して，サブバケットの転送を行う。

方法３：HCTを用いたスケジュール制御

（ｉ）　転送制御用MM（C-MM）は，自らの HCTを用いて，プロセシングクラスタへの統合及び，その転送順序のスケジュールを行う。

（ii）　C-MMは，目的のPMへの転送が可能であることを確認して，バケットの転送を開始する。

（iii）　各MMは，相対的に隣接するMM（or C-MM）の転送するサブバケット及び，転送先PMを記録しておく。目的のPMへの転送が可能になれば，そのスケジュールに合せてサブバケットの転送を行う。

これらの方法について，その相違を表１にまとめて示す。

| | TCMの資源管理 | 分配系の負荷増 | MMの負荷分散 |
|---|---|---|---|
| 方法１ | 要 | 無 | 均一 |
| 方法２ | 不要 | 有 | 均一 |
| 方法３ | 不要 | 無 | 不均一 |

表１　各スケジュール方法の相違

## 5．おわりに

バケット収集系に於ける，伝送の為のモジュール群制御の方法について，上記の３種類の方式を提案した。現在，それぞれについてのより詳細な性能の評価を行っている。詳細なスケジュールの方法を含めて，機会を改めて発表する予定である。

参考文献

［１］伏見　他，『リングバスを用いたＧＲＡＣＥのモジュール間結合系』，第29回情処全大，3F-7

［２］喜連川　他，『データベースマシンＧＲＡＣＥ＝バケットのパイプライン処理機構＝』，第26回情処全大，4F-1