3F-7

リングバスを用いたGRACEのモジュール間結合系

伏見　信也[†]　喜連川　優[††]　加藤　寿人[†††]　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　[†††]日本IBM)

1. はじめに

高並列関係データベースマシンGRACEはプロセッシングモジュール(PM)，メモリモジュール(MM)，ディスクモジュール(DM)，コントロールモジュール(CM)の4種のモジュールとそれらを結合するモジュール間結合系からなる。現在，リングバスを用いたモジュール間結合系の実装を進めており，本稿では〔1〕に引き続いてその伝送制御手順の詳細について報告する。

2. リングバスを用いたモジュール間結合系

2.1 MIMD環境

GRACEでは1つの関係代数演算処理に対し，PM，MM，及びチャネルが各々複数個割当てられる。GRACEではこの集合をタスクと呼び，CMはこれを単位としてMIMDの実行環境を実現する。従ってモジュール間結合系は複数の並列タスクが互いに干渉し合うことなく，各々自律的に動作可能な通信環境を提供しなければならない。この為に各PM，MMは自分が属するタスクの一意的なTaskIdを保持し，一方，チャネル内のヘッダ部にもチャネルが属するタスクのTaskIdを保持するTaskIdフィールドを設け，PM，MMは保持しているTaskIdと同一の値をそのTaskIdフィールドに持つチャネルのみを用いてタスク内通信を行う方式を採用した。

2.2 バケット分配とバケット収集

GRACE上では2種類のデータ流があり，その転送パタンは互いに異なる。各タスク内のPM，MM間でのデータ転送は基本的にはこれらのいづれか一方の形態をとる。

(a) PM→MM（バケット分配)　当該処理の結果タプルに対し，次処理の属性に関しhashを施し，次処理に対し割付られたMM群に向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MMにわたってできる限り均等に分散される。1つのバケットに対し，各MMに分散され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は割当てられたバケットのタプルを収集しつつ，演算処理を行う。これはPM群がパイプライン的にMMを順次訪れ，対応するサブバケットを収集することにより実現される。

(b)はPM，MMの結合が巡回的な規則性を有する為実装が比較的容易であり，以下では(a)を実現する為の伝送制御手順

3. リングバスを用いたバケット分配系

3.1 基本アルゴリズム

バケット分配の制御はMM群に分散することを基本とする。一般にタプル長はチャネルのデータ部より長く，1タプルの転送は複数のセグメント転送に分割される。セグメント転送には，(a)PM群が送出する各タプルが属するバケットに対し，リングを一周してその時点に於けるMM群内でのサブバケットサイズのMAX，MINを得るInitial Phase，(b)Initial Phaseで得られた分布情報に基づき，各MMがタプルの所属バケットに対応するサブバケットサイズBに対し評価関数R＝(MAX-B)／(B-MIN)の値が最大のタプルを求め，そのタプルを送出したPMとリンクを張るLink Phase，(c)Link Phaseで張られたリンクによってPM，MM間でタプルセグメントを転送するTransmission Phaseの3つのPhaseがあり，これらPhaseを順に状態遷移することにより1タプルの転送を完了する。タプルセグメントはLink Phase，Transmission Phaseで転送さる。Initial Phaseは前処理のPhaseであるが，これは直前に転送されるタプルの最終Transmission Phaseと重畳化出来，分布情報収集の為のオーバヘッドは実効的に存在しない。

3.2 伝送制御手順

図1にバケット分配に用いられるチャネルフォーマットを示す。前節で述べた基本アルゴリズムの実装にあたっては，各タスク内で駆動されるチャネルの個数はタスク内のMM台数に等しいものとした。一方で，一般にタスク内のPM台数nとMM台数mとの間に固定的な関係は存在せず，また全てのPMが常にタプルを送出するとは限らない。そこでPM群からMM群に対する1回のタプル群送出に対し，実際に送出されるタプルの個数をtとし，mとtとの大小関係に従

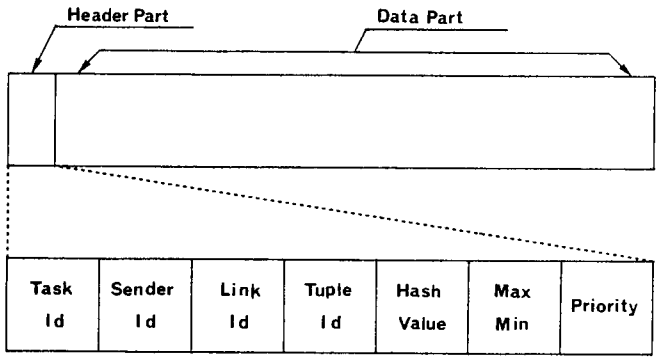

Fig.1　Channel Format

って伝送制御手順を示す。以下の手順では，各 Phaseに於てチャネルは常にPM群，MM群の順にアクセスされる様リングバス上でのモジュール群配置がされていると仮定する。

(1) t＝m の場合

〔Initial Phase〕PMは流れてくる空きチャネルを一つとり，HashValue，LinkId，及び TupleIdフィールドに各々送出タプルのHash値，自Id，タプルIdを書き込み，あわせて Max／Min フィールドを初期化してこれをMM群に向かって送出する。PMのIdとしては，当該タスクのPM群中に於ける相対位置を用いる。タプルのPM内Idは，後に述べる様に１つのPMが一度に複数個のタプルを送出する場合がある為，これらを識別する為に必要とされる。MMは送られてくるチャネル各々に対し，その HashValueフィールドの値を用いて対応するサブバケットサイズを計算し，チャネルの Max／Min フィールドの値を適宜変更する。

〔Link Phase〕PMはチャネル群からそのLinkIdフィールド値が自Idと等しいものを選び出し，その値をチャネル内のSenderIdフィールドにコピーする。また，タプルの第一セグメントをチャネルのデータ部に書き込む。各MMは送られてくるチャネルの HashValue，Max ／ Minフィールドによりタプルの R値を計算し，バッファ内に保持しているタプルのそれよりこの値が大きければSenderIdフィールド以外のヘッダ部を含めてバッファとチャネルの内容を交換する（Normal Mode，ここで初期化時のMM内のバッファはTask Idフィールドを除いて空，そのR値は -∞である）。この Phase が終了した時点で各MMにはR値最大のタプルの第一セグメント，及びそのタプルを送出したPMId，タプルのPM内Idが保持されている。

〔Transmission Phase〕各PMは，送られてくるチャネルのSenderIdフィールドと自Idを比較し，使用すべきチャネルを選択して残りのセグメントをチャネルデータ部にのせる。MMはチャネルのSenderId,TupleIdフィールドとLink Phaseで得た相手PMのId，タプルIdとを比較して所要のチャネルを選択し，チャネルデータ部のセグメントを取り込む。

(2) t＜m の場合

この場合，タプルを送出しないPMは単に１タプル転送期間中に転送動作を行わないだけでよい。一方，MMに関しては，(1)の場合の手順をそのまま実行すると，データ流方向に対して先頭のｔ台のみが有効チャネルを獲得することになる。そこで先行するｍ－ｔ台のMMは，R＞1のタプルに限って(1)のタプルの取込み動作を行うものとし（Reduced Mode），この場合に対してもタプルの均一分配を保証する方式を考案した。実際には事前にｔの値を得ることは困難なので，MMはLink Phaseに於て空チャネルの数をカウントし，これとPMと同様にして与えられた相対位置との比較を行うことによって動的にNormal Mode からReduced Modeに変化する。MMに対する具体的な手順は(1)のそれに以下の変更を加える。即ち，InitialPhase では送られてくるチャネルの内，空でないチャネルに対してのみ Max／Min フィールドの値を更新する。またLinkPhase では転送に先立ってMMは全てNormal Mode に初期化される。MMは通過した空チャネルの数をカウントしており，その値が自相対位置と等しくなった場合はReduced Modeに移行し，R＞1のタプルに対してのみ動作を行う。

(3) t＞m の場合

この場合，タスク内の駆動チャネル数はｍと等しい為，MMに対しては(1)の場合の制御で十分である。一方，PMに関しては，MMの処理能力を上回るタプルを送出することとなり，PM内での送出不能タプルのバッファリングが必要となる。この為，PMは数タプル分のバッファを持ち，バッファリングされているタプル数に応じたpriorityが与えれる。最小priorityは(2)に於ける送出タプル無しの場合に相当する。PMは自priorityとチャネルヘッダ部のpriorityフィールドを用いてInitial Phase に於て以下の様にしてチャネル獲得を競い合う。即ち，PMは送られてくる全てのチャネルに対し，自priorityとチャネルのpriorityフィールドの値を比較し，自がチャネルのそれよりも大きい場合にのみチャネルのLinkId，TupleId フィールドに対応する値を上書きし，セグメントをチャネルデータ部に書き込む。

3.3 転送の中断を許す伝送制御

何等かの理由によりPM，MMがタプルを送信，受信できなくなった場合に対しては，以下の様な手順をとる（但し，各モジュールのインタフェース部はactiveであるものとする）。PMが送信不能の場合は，そのようなPMが単にチャネルの争奪に加わらないことで十分である。一方，MMが受信不能となった場合には次の様な手順をとる。まず受信不能なMMは InitialPhase に於て適当なチャネルを選び，そのpriorityフィールドに最大値を書き込む（deadチャネル）。PMはLink Phaseで先に自分が獲得したチャネルがdeadならば当該チャネルのLinkId，TupleId，SenderId を無効化し，タプルの送出を中止する。一方，MMのLink Phaseでは，受信不能のMMはdeadチャネルに対するR値が最大となる様に，その他のMMのそれは最小となるように制御を行い，受信不能のMMを仮想的にdeadチャネルに接続することによって，受信可能なMMのみに対するタプルの均等分配が実現される。

4．おわりに

現在，上記伝送制御手順を実現するリングバスインタフェースユニットのハードウェア化を進めている。

参考文献 〔1〕喜連川，『並列処理データベースマシンに於けるデータ流制御機構』，第28回情処全大，2E-1