1B-8

GRACEプロトタイプシステムにおける プロセッシングモジュールの設計

鈴木 孝† 伏見 信也†† 喜連川 優††† 田中 英彦†† 元岡 達††

(†三菱電機 ††東京大学工学部 †††東京大学生産技術研究所)

1．はじめに

関係データベースマシンGRACE は、MIMD制御のマルチプロセッサで構成される。現在GRACE のプロトタイプシステムとしてDisk Module、Memory Module、Processing Module、Working Disk Module、Control Module各モジュールから構成されるデータベースマシンを試作中である。

各モジュールは既存の汎用計算機をホストマシンとして用い、ソフトウェアによる実行制御を行う。今回ハードウェアソータを中心に構成されるプロセッシングモジュール(PM)を新たに設計した。PMはホストマシンから転送されてくるデータストリームに沿って各種演算を実行する機能を持ち、高スループットでデータ処理を行えるよう設計されている。

2．GRACE プロトタイプ構成

Fig.1 にGRACE プロトタイプの論理的構成を示し〔1〕、Fig.2 にハードウェア構成図を示す。GRACE プロトタイプは、汎用計算機(MELCOM 80/500) 上に、各種ソフトウェアプロセス〔2〕と専用ハードウェアを用いて実現される。

次に各モジュールの機能について説明する。

(1) Disk Module

データベース格納Diskと、Selection、Projection、さらにHashing を実行するCPU 内ソフトウェアプロセスから成る。

(2) Memory Module

主記憶上に生成されるCPU によるソフトウェアプロセス。各モジュールとのデータ転送等を制御する。

(3) Processing Module

ハードウェアソータを中心とする専用ハードウェアとその入出力を制御するソフトウェアプロセスから成る。

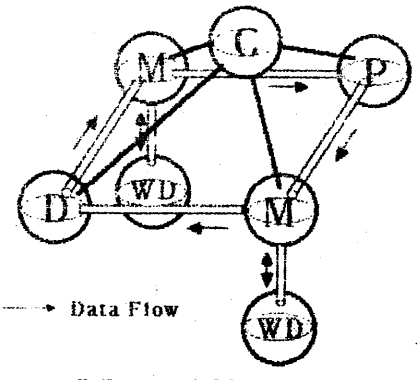
入力データストリームに対し関係代数演算等を実行する。(以下、本ハードウェアをPMと呼ぶ。)

(4) Working Disk Module

データストリームの仮想記憶化を実現するモジュール。作業用Diskと仮想化を行うソフトウェアプロセスから成る。

(5) Control Module

各プロセスを制御するホスト上のソフトウェアプロセス。



C:Control Module
M:Memory Module
P:Processing Module
D:Disk Module
WD:Working Disk Module

Fig.1 GRACE Prototype Logical Structure

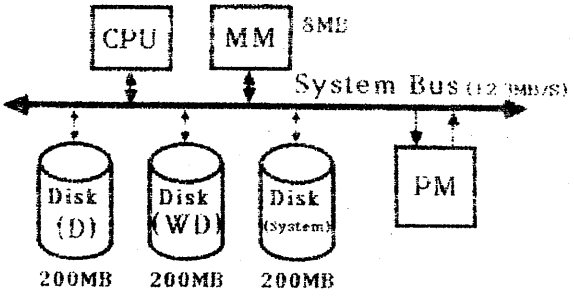


Fig.2 GRACE Prototype Hardware Configuration

3．プロセッシングモジュール

3－1．内部構成

Fig.3 にプロセッシングモジュール(PM)の構成図を示す。PMは関係代数演算処理におけるSort、Join、Aggregation 等の演算を行うモジュールである。PMはハードウェアソータから成るソートユニットを中心に構成され、さらに周辺ユニットとして、ソートユニットの制御系や出力データストリーム処理、すなわち各種演算を実行するユニットが必要となってくる。

PMは次の各ユニットから構成される。

(1) System Bus Interface

システムバスに接続され、PMの制御情報の受信、リレーションのタプルデータ入力、ソート結果の出力、関係代数演算後のタプルデータの出力を行う。また入力ポートと出力ポートが独立しており、入出力同時動作が可能となっている。このユニットにはデータ入出力ポートの他に68000との制御情報の送受を行う入出力ポートも接続され、68120 (Intelligent Peripheral Controller) のDual Port RAM を介して68000 I/O Channel と通信を行う。

(2) Sort Driver 〔3〕

ソートユニットを制御するハードウェアドライバで、入力レコードへのフラグ生成付加等をデータストリームに沿って実行し、ソートユニットに送出する。さらにソート結果を次段のユニットに送出する。

(3) Sort Unit

パイプラインマージソートを行うハードウェアソータから構成され、大容量のソートを行うことができる。

(4) 68000 Computer System

ソートされたタプルに対し、JoinやAggregation 演算を行い、次JoinのためのHashing を行うユニット。演算は68000 のソフトウェアで実行される。また68000 は、これらの演算とデータ入出力のDMA 制御をリアルタイムOSのマルチタスク機能により管理する。さらにシステムバスからのPM制御情報を基にPMの各ユニットを初期化する機能も有する。

3－2．モジュール／ユニット間インターフェイス

GRACE の各種演算は、データストリームが一巡する間に処理が施されていくデータ流指向の処理方式を取っている。プロトタイプ試作に当たってPMを設計する際にも、データストリームに沿って処理が順次進行できるよう、特に各ユニット間データ転送を効率良く行うインターフェイスを考慮した。すなわち各ユニット間のデータ転送中に存在するオーバーヘッドをできるだけ排除し、高スループットを実現するハードウェアを提供することにより、データ流に沿った効率の良い処理が実現できる。

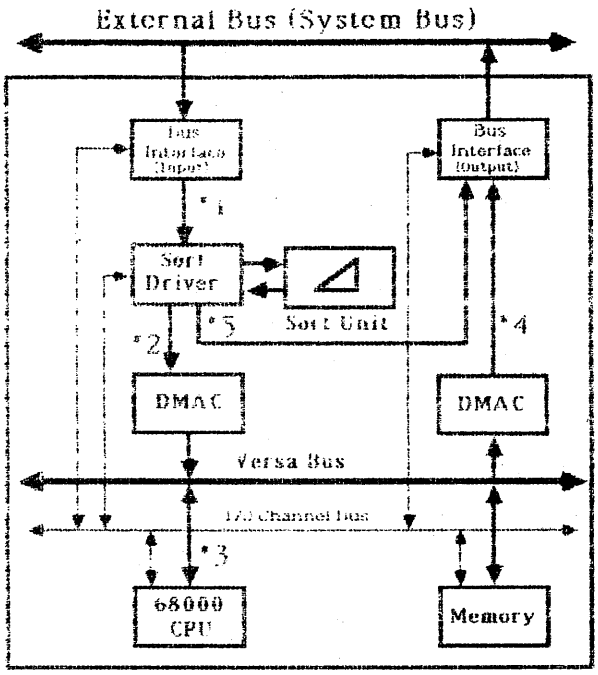


Fig.3 Processing Module

PM内部構成は、GRACE におけるプロセッシングモジュールとして汎用的なアーキテクチュアとなっている。今回プロトタイプとして汎用計算機との結合を考えるにあたり、外部バスインターフェイスとして最も高速なシステムバスインターフェイスを新たに設計した。この部分は計算機システムに強く依存するが、データベースマシン全体のスループットを考えた場合大きな要素を持ってくる。このためPMはDiskと同様にOS中のI/O ハンドラの直接の制御を受けることのできるプログラムインターフェイスと、主記憶、PM間の高速データ転送のためのハードウェアインターフェイスを持つ設計とした。

4．データストリーム転送制御

4－1．処理の流れ

PM内のデータ処理の流れを以下説明する。

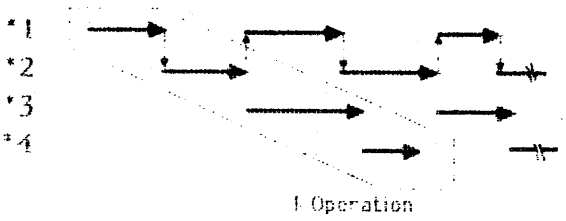
ホスト上のMemory Module を実現しているプロセス（Mプロセス）からデータ転送に先立ち送出されてくるPM初期化命令を受け取ったシステムバスインターフェイスは、その内容、すなわちレコード長、レコード数、Key 情報、Aggregation 演算情報等を68000 に送出する。68000 はその内容を基にして、Sort Driver 、Sort Unit を初期化する。それが終了すると、以後のレコードデータは、Mプロセスからシステムバスを経由して、Bus Interface、Sort Drive、Sort Unitの順にデータストリームの形で流れる（＊１の経路）。この間は3MB/s 程度のスループットで処理が進む。ソートされたレコードは、引き続き、Sort Driver を経由して次のユニットに送られる。ここで処理は二つに分れる。ソート処理のみの場合は＊５を経由し、Bus Interface(Output) を通し、システムバスを介して直接Mプロセスに答えが返される。この経路はすべてハードウェアによる転送制御を用いており、Mプロセスはデータ出力後ただちにソート結果を受け取ることができる。JoinやAggregation 演算を行う場合は＊２を経由し、68000 のDMA Controllerを介して一旦68000 メモリ空間に入れられる。68000 はソフトウェア処理により（＊３）、ソートされたレコードからJoinしたタプルの生成、あるいはAggregation演算を行い、その結果が＊４を経由し、システムバスを介し、Mプロセスに目的の答えとして返される。

4－2．データ入出力／演算パイプライン

Fig.4 にデータ入出力／68000 演算のパイプライン制御を示す。

本PMは、システムバスインターフェイスとして、入力ポートと出力ポートを独立に持っており、Join演算（＊３）と次オペーレーションのデータ入力（＊１）あるいはソート出力（＊２）、Joinタプル出力（＊４）と＊１あるい＊２がパイプライン処理できる。

それぞれのプロセスの処理時間は一定ではないため、各プロセス間の同期化が必要となってくる。現在ではJoin演算は68000 ソフトウェアで行うためタプル数が多い場合は時間を要し、理想的なパイプラインとはならず、スループットが落ちることも考えられる。68000 の処理時間を含めたシステム全体のスループットについては今後の評価検討を要する問題である。



*1:Data Stream Input
*2:Sorted Data Stream Output
*3:68000 Join Process
*4:Result Output

Fig.4 Data Transfer / Join Process Pipeline Flow

5．おわりに

GRACE プロトタイプにおけるプロセッシングモジュールの設計により関係データベースにおける負荷の重いJoin等の演算を高速に行うことが可能となる。

現在、各ユニットの詳細論理設計を進めている。

今後の課題として、データ転送に伴うオーバーヘッドの評価、Joinアルゴリズムの効率良い方法とそのハードウェア化等について検討する予定である。

＜参考文献＞

［1］ 伏見、喜連川、田中、元岡：
「データベースマシンGRACE のプロトタイプシステム」
情報処理学会第３１回全国大会 １Ｂ－６、１９８５

［2］ 中山、伏見、喜連川、田中、元岡：
「GRACE プロトタイプシステムにおけるソフトウェア構成」
情報処理学会第３１回全国大会 １Ｂ－７、１９８５

［3］ 鈴木、伏見、喜連川、田中、元岡：
「ハードウェアマージソータの駆動系の設計」
情報処理学会第３０回全国大会 １Ｄ－９、１９８５