6L-8

コネクションマシンCM-2における関係データベース処理

松本 和彦　喜連川 優　高木 幹雄
東京大学 生産技術研究所

1 CM-2アーキテクチャ

コネクションマシンは、数万個というオーダーのプロセッサアレイとハイパーキューブ形状の通信ネットワークを持つ、単一命令ストリーム多重データ(SIMD)の並列アーキテクチャマシンである。現在Thinking Machines Corporationが市販している最新モデルはCM-2と呼ばれ、フルセットで65536個のプロセッサを備えている。

CM-2は、ワークステーションなどのフロントエンドにバスインタフェースを介して結合している。逐次的な処理はフロントエンド上で通常のプログラムコードとして実行され、並列処理の部分だけがバスインタフェースを通じてコネクションマシンに指令される。

CM-2の操作は、PARISというライブラリの関数を呼び出すことによって行なう。次節で、コネクションマシンのいくつかの基本概念について述べる。

2 PARISインタフェース

PARISとは、コネクションマシンシステムをプログラミングするための、PARallel Instruction Setのことである。それは、フロントエンドからコネクションマシンプロセッサを指令するための低レベルインターフェースである。PARISの実体は、関数、マクロ、変数の集合である。関数やマクロは、呼び出されることにより、必要に応じてコネクションマシンのシーケンサに命令を送出する。

2.1 仮想プロセッサ

PARISは、抽象化されたコネクションマシンハードウェアをユーザに見せることによって、スケーラビリティを拡張している。その核心となるのは、各々の物理プロセッサが複数の仮想的なプロセッサをシミュレートするという、仮想プロセッサの機能である。

プログラマは好きなだけの仮想プロセッサ数を仮定してプログラムすれば良く、プログラムのできた後で、仮想プロセッサは物理プロセッサに割り当てられる。こうすることによって、ソフトウェアが物理的なハードウェアによって変更を強いられることは無くなる。そこに残るのは、単に、物理プロセッサ数と実行時間やメモリ容量との間のトレードオフである。また、1つの物理プロセッサがいくつの仮想プロセッサをシミュレートするかという数のことをVP比と呼ぶ。

2.2 VPセット

あるデータ集合が割り当てられている仮想プロセッサの集合のことを、仮想プロセッサセット(VPセット)と呼ぶ。仕事によっては、複数のデータ集合を扱うことが必要となるので、PARISは複数のVPセットが同時に存在することを許している。例えば、ドキュメント検索のプログラムでは、ある時点では記事(数キロバイト)をデータ要素として扱い、またある時点では記事中の単語(数十バイト)をデータ要素として扱う。このプログラムでは、それぞれに対応した2つのVPセットを使う。単語をデータ要素とするVPセットは、記事をデータ要素とするVPセットよりはるかに大きくなる。VPセットはPARISへの関数呼び出しによって生成される。その大きさ(VPの数)は生成された時点で決まり、以後変化しない。VPの数は物理プロセッサの倍数でなければならない。

2.3 フィールド

メモリは、フィールドという単位で扱われる。フィールドとは、メモリ中に割り当てられた連続するビット群のことである。フィールドの長さは任意で、割り当てらる時にユーザが指定する。フィールドは、必ずどれか一つのVPセットに所属する。フィールドをVPセットに割り当てるというのは、VPセット中の全てのVPについて1つずつ、メモリ中の同じ位置に記憶を割り当てるということに相当する。

2.4 コンテクストフラグ

PARISの仮想プロセッサの持ついくつかの1ビットフラグのうち、特に重要なのはコンテクストフラグである。コンテクストフラグが立っている仮想プロセッサはアクティブであると言われ、ほとんどのPARISインストラクションはアクティブなプロセッサでのみ実行される。

3 コネクションマシン上でのジョイン処理

コネクションマシン上のアプリケーションには、有限要素解析等の数値計算、コンピュータビジョン、意味ネットワークなど、様々な分野のものが考えられているが、非常に多くのデータアイテムに似通ったオペレーションを行なうデータベース処理は、特に適したアプリケーションであると思われる。

CM-2上で実際にジョイン処理を行なうプログラムを作成し、その性能評価を行なった。ジョイン処理(結合演算)は、関係データベース処理の中心的なオペレーションであり、その手法には様々なものがあるが、今回コネクションマシン上にインプリメントしたものは、マージに基づくアルゴリズムである。

4 アルゴリズム

ジョインの手順を以下に示す。

1. 2つのVPセットA、Bを作り、それぞれに元となるリレーションをロードする(図1)。
2. scan-with-addオペレーションを使い、それぞれのプロセッサが、自分と同じキーを持つプロセッサ(タプル)が自分よりも上方にいくつあるのかを調べる(図2)。

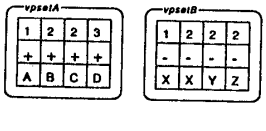

図1

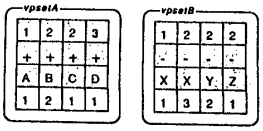

図2

3. それぞれのキーの先頭にあたるプロセッサを駆動し、キーの値を宛先アドレスとして自分と同じキーのタプルの数を新しいVPセットLに送信する(図3)。

---

[0] Relational Database Processing on the Connection Machine CM-2
K.Matsumoto, M.Kitsuregawa, M.Takagi
The Institute of Industorial Science, University of Tokyo

4．Lでは、メッセージを送ってきたタプルがマージにより移動すべきアドレスを計算し、そのアドレスをそれぞれのプロセッサに送り返す（図4）。

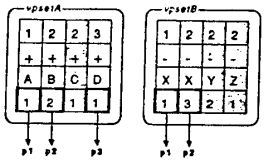

図3

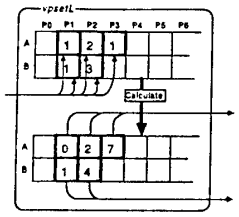

図4

5．AとBでは、返ってきたアドレスをもとに、それぞれのキーの先頭以外のタプルが移動すべきアドレスをセットアップする。

6．AとBの全てのタプルが、新しいVPセットMの中の、移動すべきプロセッサへと送信される。この時点でマージは完了する。

7．それぞれのキーにおいて、Aから送られてきたタプル数とBから送られてきたタプル数を数え、それを掛け合わせる（図5）。

8．7.で求めた数を使い、scan-with-addオペレーションによって、ジョイン後にそれぞれのキーの先頭タプルが位置するアドレスを計算する（図6）。

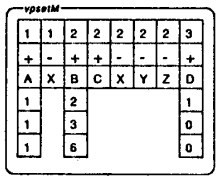

図5

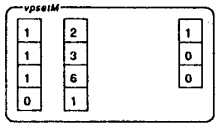

図6

9．8.の計算結果をもとに、キーと属性を新しいVPセットJの適切なアドレスに送信する（図7）。

10．scan-with-copyで、必要なだけのデータを複製する（図8）。

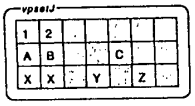

図7

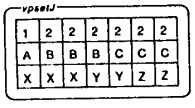

図8

11．一番下のフィールドで置換を行ない、ジョイン後のリレーションをJ内に完成させる（図9）。

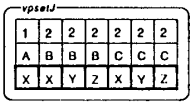

図9

5 性能評価

選択率100%のジョインの場合の性能測定の結果を、表1に示す。タプル長はキー32ビットと属性32ビットの計64ビットである。なお、この時間にはリレーションの入出力の時間は入っておらず、純粋にプロセッサがマージとジョインを行なっている時間のみが含まれる。

タプル数	処理時間	VP比	タプル数	処理時間	VP比
100	56.5	1	10000	267.5	4
200	67.7	1	12000	215.0	4
500	81.8	1	14000	227.7	4
1000	71.5	1	16000	244.0	4
2000	74.0	1	20000	334.5	8
3000	80.3	1	24000	339.2	8
4000	97.0	1	28000	361.2	8
5000	123.7	2	32000	368.7	8
6000	121.2	2			
7000	145.5	2			
8000	145.3	2			

表1: ジョイン処理の時間（msec）

6 終りに

高並列微粒度マシンCM-2にメインメモリデータベースを実現し、その基本性能を測定した。演算は全て主記憶上でなされており、入出力は含まれていない。高速ディスクDataVaultを使ったより大規模なデータベース処理が、今後の課題である。

参考文献

[1] W.Daniel Hillis, "The Connection Machine",The MIT Press,1985

[2] Thinking Machines Corporation, "Connection Machine Technical Summary",1989