知識ベースマシンＭｕ－Ｘ(2)
―並列処理のための基本機能―

酒井　浩，　柴山　茂樹，　仲瀬　明彦　　伊藤　文英
（株）東芝　総合研究所　　ＩＣＯＴ

１　はじめに

われわれは通産省第５世代コンピュータプロジェクトの一環として，項を基本要素とする非正規型関係モデルを採用した知識ベースマシン[1] を試作中である。本マシンは図１に示す構成を有するマルチプロセッサシステムであり，共有記憶として通常の共有メモリとマルチポート・ページメモリ（ＭＰＰＭ）[2] を備えて，要素プロセッサ（ＰＥ）台数が数～１００程度の並列処理システムを指向している。

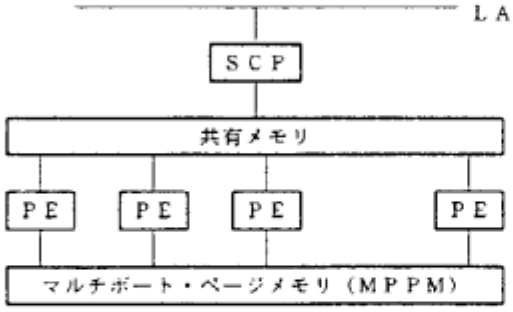
本稿では，このハードウェア上に試作を進めている知識ベース管理システムについて，処理方式の概要と，それを実現するためのＰＥ間通信機能とデータ管理機能について述べる。

２　知識ベース処理方式の概要

２．１　知識ベース並列処理方式

一般に並列処理システムにおけるＰＥ間の通信には，共有メモリを用いた比較的小規模な並列を指向する方式と，物理的にメッセージを送受し高並列を指向する方式がある。本システムは，基本的に前者に属するが，ＭＰＰＭという比較的大きなバンド幅の共有記憶を併用することにより，共有メモリのみのシステムと比較して，より高並列が実現できると考えている。

本システムにおける並列処理では，リレーションデータに対する関係代数レベルの演算を対象とし，各時点であいているＰＥが，処理すべき演算要求を捜して自発的に処理する方式を採用した。これは共有メモリによる通信の長所を生かし，ＰＥ間の負荷分散を僅かなオーバヘッドで実現することをねらったものである。



ＳＣＰ　　：MC68010，MM：2MB，DISK：40MB
ＰＥ×８　：MC68020，MM：2MB，DISK：40MB

図１　ハードウェア構成

２．２　リレーションの管理・格納方式

本システムでは，リレーションをトランザクション中でのみ有効なテンポラリリレーションと，トランザクションを越えて存在するパーマネントリレーションに区別した。そしてテンポラリリレーションは高速アクセス可能な共有記憶であるＭＰＰＭに格納し，複数ＰＥによる並列処理による高速化をはかった。一方，パーマメントリレーションは複数ＰＥへの分散配置（デクラスタリング）による並列処理と静的クラスタリング，２次インデクスによる高速化をはかった。デクラスタリング方式としては，とりあえずラウンド・ロビン法（リレーションの各ページを各ＰＥに順番に格納する）とハッシュ法（タプルをある属性でハッシュした結果で格納するＰＥを決める）の２種類を実現し，比較評価することにした。

リレーションの排他制御の単位は，リレーションおよびページの２種類とした。

２．３　基本ソフトウェアの考え方

一般にマルチプロセッサシステムでは，単一プロセッサシステムの場合と比較して同期処理等を行なう必要を生じ，一般にソフトウェアが複雑することが多い。しかしながら，本システムでは各ＰＥに独立性の高い仕事を割当てることによりソフトウェアの簡単化をはかることにした。即ち，複数のホストから受信した種々のキュエリを並行して処理する場合，単一プロセッサシステムでは各ホストに対応してタスクを割当てることが多いが，タスク切換えの処理とタスクの優先度の制御が困難になる。そこで，本システムでは，各ホストをいずれかのプロセッサに割当ててタスク切換えが本質的に起きないようにした。ただし，このため同時に接続可能なホストの数は，システムのプロセッサの台数で制限されることになる。

また，本システムでは図２に示すように，各ＰＥでキュ

Knowledge Base Machine Mu-X
Hiroshi SAKAI　Shigeki SHIBAYAMA　Akiniko NAKASE　Toshiba Corporation
Fumihide ITOH　　Institute for New Generation Computer Technology

エリ解析・実行制御と関係代数レベルの演算の並列処理の両方を受け持たせることにしているので，両者の優先度が問題となる。これについては，代替プロセッサのないキュエリの解析・実行制御を常時実行し，処理の区切りで並列演算部に制御を渡す方式とした。これにより，プロセッサやメモリなどの有効利用をはかっているが，その妥当性については今後評価を行なう必要があると考えている。

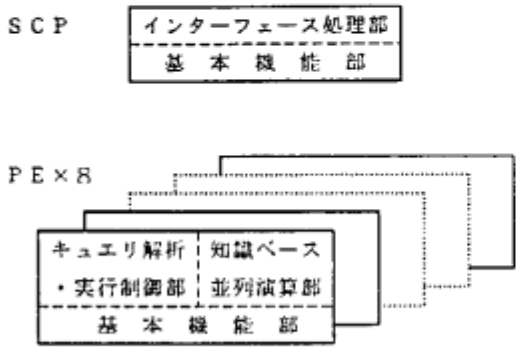


図２　ソフトウェア構成

3　プロセッサ間通信機能

本機能は，(1) フロントエンドプロセッサの役割を担うＳＣＰと実質的な処理を行なうＰＥとの間で行なうホストとのキュエリおよび処理結果の送受信，(2) ＰＥ間で行なわれる知識ベース演算の並列処理コマンドとそれに対するレスポンスの送受信に使用する。

通信機能を実現するため論理的な通信チャネルを設けた。各通信チャネルは１つのコマンド発行プログラムと複数のコマンド受信・実行プログラムを想定している。(1) ではＳＣＰのインターフェース処理部とＰＥのキュエリ解析・実行制御部の間で通信用チャネルを１つずつ使用し，(2) では各ＰＥのキュエリ解析・実行制御部とすべてのＰＥの並列演算部の間の通信に１つずつ使用する。

通信チャネル経由で送られるコマンドには，受信すべきＰＥを指定したものと，ＰＥを指定せず，同時に処理可能なＰＥ台数の上限値だけを指定したものがある。前者は各ＰＥが個別に持つディスク装置内のデータに対する処理に，また後者はＭＰＰＭ上に存在し，どのＰＥからもアクセス可能なデータに対する処理に対応する。本システムでは，共有メモリに格納されたひとつのコマンドを複数のＰＥにより並列処理することになっており，これによりコマンドコピーの処理をせずにすみ，また並列処理に必要なＰＥ間の同期処理にもそのコマンドの領域を使用することで並列処理のオーバヘッドを小さくしている。

ただし，本機能においてはひとつのトランザクションはコマンドを１つずつ逐次的に並列処理することしか実現しておらず，本機能の妥当性について今後評価をする必要があると考えている。

4　データ管理機能

本機能はリレーションをファイルとして扱う。ファイルには，ディスク装置に格納される各ＰＥごとにローカルなものと，ＭＰＰＭに格納される各ＰＥからアクセス可能なものの２種類がある。

(1) 高速ランダムアクセス

本システムでは静的クラスタリングとそれに基づくアクセスも知識ベース演算の一部と考え，他の処理とオーバラップすることにより高速化をねらった。そこで新たに作成するデータ管理機能ではページを単位とするランダムアクセスの高速化のみを実現し，インデックス処理などは並列演算部にさせることにした。

また，並列処理方式の都合により，ＭＰＰＭ上のリレーションに対するページの追加では，最終ページの直後だけでなく離れた位置に直接に追加できる機能を実現した。

(2) 排他制御

本システムでは，リレーションの排他制御の単位としてリレーション全体およびページの２種類を用意することにした。このうちリレーションの排他制御はデータ辞書中で行うことにし，データ管理機能ではページに対する排他制御を実現した。

(3) 更新ページの扱い

ファイルの更新は，更新後のページデータをＭＰＰＭに格納することにより行う。これにより，コミットとロールバック処理の高速化と，更新ページに対する高速アクセス（ディスク装置上のファイルの場合）が実現できた。

(4) 多数ファイルのサポート

結合演算や集合演算の並列処理に，喜連川らのGrace で採用されたハッシュ[3] を用いる方法を多用することにした。このため多数のファイルを実現するとともに，それらを即座に確保・解放する機能を実現した。

5　おわりに

マルチプロセッサ構成の知識ベースマシンの処理方式の概要とそれを実現するための基本機能について述べた。マシン・アーキテクチャを生かし処理オーバヘッドの少ない処理方式をねらっているが，そのために削った機能もあり，その是否についてはシステムの性能を実測することで評価したいと考えている。

参考文献

[1] 柴山他，" 大規模知識ベースマシンの開発(2)"，情報処理学会第34回全国大会論文集

[2]Tanaka, Y.,"A Multiport page-memory Architecture and a Multiport Disk-Cache System", New Generation Computing, 2, Ohmusha, 1984.

[3] 中野他，" 密結合マルチプロセッサにおける関係代数演算の評価"，情報処理学会第35回全国大会論文集