3P−5

階層記憶型並列処理の制御ソフトウェア

加藤　喜郎，橋本　伸，本間　義夫，岡田　信，神谷　幸男
富士通（株）

1．はじめに

科学技術計算分野への適用を目的とする，階層記憶型の並列処理システム（ＨＰＰ）[1]を開発した．共用記憶装置を用いた階層記憶型のシステムでは，主記憶共用の密結合型と異なり，並列処理制御及びデータ転送のオーバヘッドが大きくなる．本稿では，ＨＰＰシステムの制御ソフトウェアについて，効率のよい並列処理方式と共用記憶管理方式，及びユーザ支援ツールとの連携[2]について報告する．

2．制御ソフトウェアの機能

本稿で述べる制御ソフトウェアは，図1に示すＨＰＰシステム上で，科学技術計算を高速に行うことを目的としており，以下の機能を持つ．

－　並列実行制御機能
　並列実行単位（プロセス）の管理，同期機構など．
－　階層記憶管理機能
　ＣＳＵ空間の割り当て，ＰＥＳ－ＣＳＵデータ転送など．
－　支援ツール連携機能
　ユーザにチューニング情報などの提供を行う，ユーザ支援ツールのためのトレース情報（ログ）出力機能．

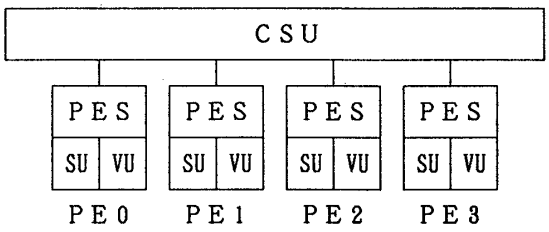

図1　ＨＰＰシステム構成図

3．並列処理方式

ＨＰＰシステムでは，ＰＥ上で動作する“タスク”が仮想プロセッサとしての働きをする．

タスクは，並列実行の単位であるプロセスを実行する．Phil言語でクラスタとして記述された手続きの実行がプロセスである．同一のクラスタが2度実行されたとき，制御ソフトウェアはそれらを2個のプロセスとして管理する．

タスクとプロセスとの割り当て処理方式は，キューにつながれたプロセスを実行待ち状態のタスクがキューから外して実行する，ダイナミックなスケジューリング方式である．この方式を採ると，ＰＥの性能が均等でない場合でも実行状況に合わせて適当な負荷分散となることが実証できた．図2に，子プロセスの生成と実行開始の処理を表す．

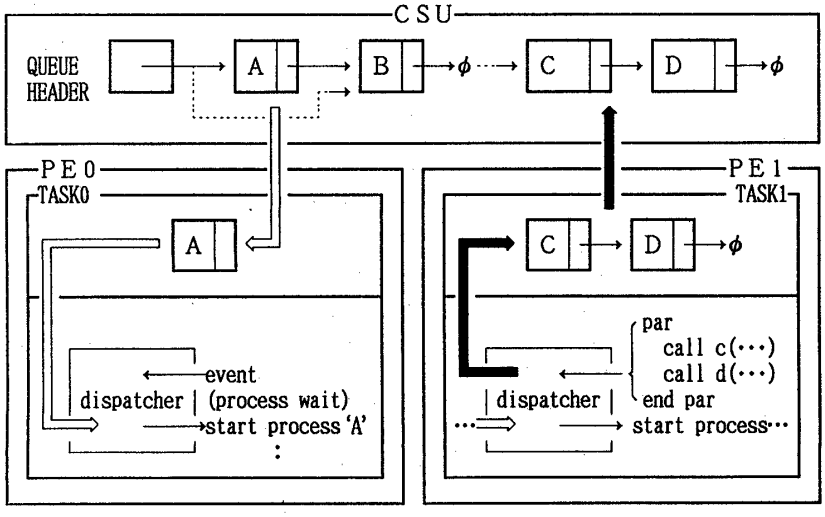

図2　プロセスの生成と実行開始

1）プロセス生成

プロセスを新たに生成するプロセスを親プロセス，生成されるプロセスを子プロセスと呼ぶ．また，同時に生成された子プロセスを兄弟プロセスと呼ぶ．

子プロセスを生成するときは，子プロセスの入口点，パラメタなどの情報をプロセス制御ブロック（ＰＣＢ）に格納する．ＰＣＢをＣＳＵ上のキューに接続することによって，その子プロセスは実行待ち状態となる．

親プロセスは，子プロセスを生成すると全ての子プロセスが終了するまで待ち状態となる．親プロセスを実行していたタスクは，他のプロセスを実行することができる．

The Parallel Processing Control Program for Hierarchical-memory Connected Multi-processor

Yoshiro KATO, Shin HASHIMOTO, Yoshio HONMA, Shin OKADA, Sachio KAMIYA
FUJITSU, Ltd.

2）同期機構

Phil言語で提供される，post文，wait文，lock文，barrier 文に対応したライブラリで同期機構を実現した．

post／wait及びlockでは，ユーザがＣＳＵ上に宣言したイベント変数及びロック変数をキューヘッダとして，排他制御と待ち状態となったプロセスのキューイングを行う．

barrier では，制御ソフトウェアがbarrier 変数をＣＳＵ上に確保する．barrier 変数にはキューヘッダの他に子プロセス数を格納するフィールドを持ち，兄弟プロセスの待ち合わせに使用する．

3）入出力機能

入出力は，ＰＥ０で動作する特定のタスク（TASK-0）で全て動作する．この機能のためＰＥ台数が増えても，入出力装置が１組あれば並列ジョブが動作することができる．

入出力に関する情報は，要求元からＣＳＵ経由でTASK-0の入出力処理ルーチンに渡る．

4．階層記憶管理

制御ソフトウェアは，ＣＳＵのアクセスを極力減らし，オーバヘッドが小さくなるよう最適化を行った．Phil言語で記述されたＣＳＵ−ＰＥＳ転送は，実行時に可能であればＰＥＳ上でバファリングして一括転送を行った．制御情報は，以下のような方法でアクセス回数の削減を行った．

− 初期化時に値が定まるテーブルはＰＥＳにコピーして参照する．

− 連続した領域になるよう割り付ける．

各プロセスが使用するＣＳＵ領域は，初期化時に各タスクに分配した領域を，各タスク専用の管理テーブルを使って管理した．管理テーブルはＰＥＳに置くことができるので，ＣＳＵ領域の獲得のためにＣＳＵをアクセスする必要がなく，排他制御のオーバヘッドも削減できた．

5．ツール連携

制御ソフトウェアには，実行時にログを出力する機能がある．ログは，子プロセス生成，ＰＥＳ−ＣＳＵデータ転送開始・終了などの事象ごとに，事象の発生時刻，転送データ量などを記録した情報である．ログはログバッファが一杯になったとき，及びジョブの終了時にログ用ファイルに出力する．ログファイルはユーザ支援ツールの入力データとなる．また，簡易なツールを用いて，実行時ライブラリ自身のデバッグにも使用した．

ログを出力すべき事象は，ＰＥＳ−ＣＳＵの転送最適化など，目的により異なる．また，デバッグが目的の場合は異常終了する直前の数レコードが得られればよいこともある．この問題解決には，チューニングツール又はデバッグツールとの，より有機的な連携が必要である．

6．成果

ＨＰＰシステムの性能評価のために，以下の評価用プログラムの並列実行を行った．

− ＳＰＩＮ：　磁場解析，転送行列法

− 円周率計算：　並列ＦＦＴによる多倍長計算

− ＮＩＥＬ：　橋の強度解析

− 帯行列解法：　疎行列係数の連立一次方程式の解法

なお，ＳＰＩＮ及び帯行列解法は，高水準並列処理記述言語ＰＡＲＡＧＲＡＭ，他はPhil言語で記述されたプログラムである．

表１は，評価プログラムの実行時間（経過時間）及び台数効果である．表２は，制御ソフトウェアのオーバヘッド及びグラニュラリティの代表的な値である．

表１　評価プログラムの実行性能
経過時間　単位：秒

プログラム	１タスク 経過時間	４タスク 経過時間	４タスク 台数効果
帯行列解法	284	100	2.8
ＳＰＩＮ	1811	479	3.8
円周率計算	3521	1375	2.6
ＮＩＥＬ	228	108	2.1

表２　オーバヘッドとグラニュラリティ

子プロセス生成	1 ms
lock文（成功）	0.7 ms
lock文（失敗，別プロセス起動）	0.8 ms
評価プログラムのグラニュラリティ	0.8〜7.8 sec

7．おわりに

実用プログラムの並列実行に成功し，ＨＰＰシステムの有効性が実証できた．評価に用いた大規模な科学技術計算プログラムでは，制御ソフトウェアのオーバヘッドは，グラニュラリティに比べて十分小さいことが分かった．

ツール連携のためのログ出力機能は，ユーザ支援のみでなく，制御ソフトウェア自身のデバッグにも有効であった．

謝辞　本研究は，通商産業省工業技術院大型プロジェクト「科学技術用高速計算システムの研究開発」の一環として，新エネルギー・産業技術総合開発機構(NEDO)から委託を受けて，実施したものである．

御協力頂いた，日本原子力研究所物理部　別役主任研究員，日本原子力研究所東海研究所計算センター　横川氏，東京大学大型計算機センター　金田助教授，京都大学工学部土木工学教室　渡邊教授に感謝する．

【参考文献】

(1) 橋本，他：階層記憶型並列処理，情報処理学会第41回全国大会，1990

(2) 渡辺，他：並列処理におけるプログラム開発支援システム，情報処理学会第41回全国大会，1990