マルチＰＳＩ第２版のハードウェア構成

武田　保孝，中島　浩，益田　兼直，瀧　和男
（三菱電機）　（ＩＣＯＴ）

１　はじめに

マルチＰＳＩシステムは，並列推論ソフトウェアの研究開発環境を提供することを目的とした並列計算機である。マルチＰＳＩの開発は第１版と第２版の２段階に分けて行われており，既に完成した第１版は６台のＰＳＩ（Personal Seqential Inference Machine）を格子型ネットワークで接続したものである。現在開発中の第２版は第１版の評価結果を基に，プロセッサ数の増加及び接続ネットワークの改良による大規模化・高性能化を行なったものである。本稿では，接続ネットワークを中心としてマルチＰＳＩ第２版の特徴とハードウェア構成について述べる。

２　マルチＰＳＩ第２版の概要

マルチＰＳＩ第２版の全体構成を図１に示す。マルチＰＳＩ第２版は，フロントエンド・プロセッサとしてＰＳＩを小形化・高性能化したＰＳＩ－Ⅱを用いたバックエンド・マシンであり，第２版本体上では並列推論言語ＫＬ－１(Kernel Language vergion 1) が実行される。

フロントエンド・プロセッサは，本体と接続するネットワークパス及びメンテナンスパスを介して入出力・デバッグ・保守などを行なう。本体は，６４台のＰＥ(Processer Element) を接続用ハードウェアを含して８×８の格子型ネットワークに接続したものであり，各ＰＥのＣＰＵはＰＳＩ－ⅡのＣＰＵとほぼ同じものである。ＰＥ間の通信を高速に行なうために，全ＰＥが同期クロックで動作する。

マルチＰＳＩ第２版の特徴として，P^3 (Processing Power Plane)を用いた負荷分散方式があげられる。P^3 は処理能力が均一に分布するような座標平面であり，この平面を適宜分割しその各々をゴールに割り当てることにより静的負荷分散を行なう。図２の例では，ｐに割り付けられたP^3 領域をｑとｒで等分し，ｒに割り付けられた領域をさらに３等分している。物理的なプロセッサとの対応付けは，P^3 をプロセッサ数に分割し，各ＰＥが自身に割り付けられた領域上のゴールを実行することにより行なう（図３－１）。計算過程においてＰＥ間で負荷の不均衡が生じた場合，そのＰＥ間の境界を動かすことにより受け持ち領域の面積を変えて負荷の均衡を図る（図３－２）。この方式は，局所的に負荷の調整を行なえるため，大規模な並列計算機システムに適した方法と考えられる。

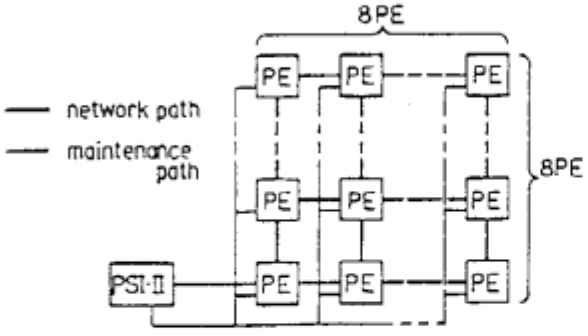


図１　マルチＰＳＩ第２版の全体構成

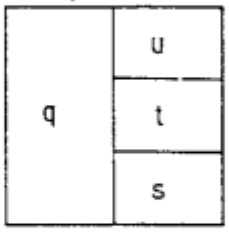


図２　ゴールへのP^3 の割り付け

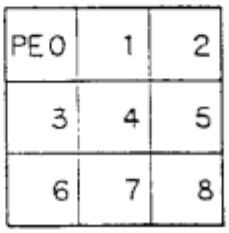


図3-1 静的負荷分散

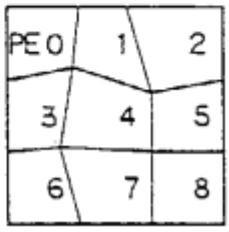
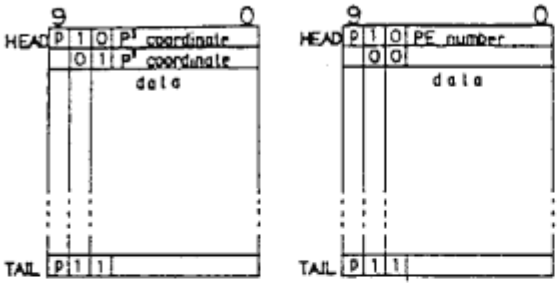


図3-2 動的再割り付け

３　パケットのルーチング方式

マルチＰＳＩ第２版では，動的な負荷分散を行なうためにP^3 に於ける各ＰＥの受け持ち領域は動的に変化する。そのため，個々のＰＥが全体の処理の分配状況を知ることは困難であり，かつ処理の局所性を考えると好ましくない。そこで，図４に示すように，ＰＥ間通信に用いるパケットを，その宛て先がP^3 の座標で示されたものと（ａ），物理的なＰＥ番号で指示されたもの（ｂ）の２種類設け，ゴールの実行を要求するパケットは（ａ）を用いることとした。

Hardware Configuration of the Multi-PSI/Version-2
Yasutaka Takeda, Hiroshi Nakashima, Kanae Masuda (Mitsubishi Electric Corp.)
Kazuo Taki (ICOT)



（ａ）P3 座標表示　　（ｂ）ＰＥ番号表示

図４　パケットの構造

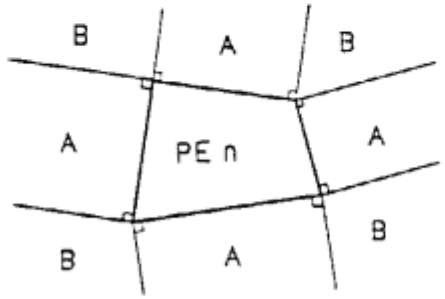


図５　パケットの転送方向決定法

このようなパケットの転送時には，パケットの通過経路上にあるＰＥも全体の状況を知らないため，パケットの転送方向決定はなんらかの局所的な情報で行なう必要がある。そこで，図５のように各ＰＥがP^3上の受け持ち領域の各頂点から辺に対する垂線を立て，パケットの目標座標が辺に垂直な区域Ａに属するときはその辺を共有するＰＥにパケットを転送し，区域Ｂに属するときはその頂点を挟む両側の辺を共有する２つのＰＥのうちどちらか一方にパケットを転送することとした。このようにすれば転送経路上のＰＥの受け持ち領域とパケットの目標座標の距離は増加することがなく，パケットは必ず目標座標に対応するＰＥに到着する。従って，各ＰＥはP^3における自身の受け持ち領域情報だけから，パケットの正しい転送方向を求めることができる。

4　ハードウェア構成

図６に接続用ハードウェアの構成を示す。接続用ハードウェアはＰＥの内部バスに接続されており，その内部資源はＣＰＵの特殊レジスタとしてアクセスできる。接続用ハードウェアは，隣接するＰＥの接続用ハードウェアへの４本のチャネル（Ｃｈ０－３）とＣＰＵへのチャネル（Ｃｈ４）が５×４のスイッチで接続されている。パケットの転送は，同一チャネルへの複数の転送要求がないかぎり並列に行われる。各チャネルは，パリティを含めた10ビット単位で転送を行ない，１チャネルの１方向当たりの転送速度は５Ｍバイト／秒である。

Ｃｈ０－３には，P^3の座標またはＰＥ番号に対応した転送方向を格納するＰＴ(Path Table)がチャネル毎に設けられてあり，パケットの到着時にその宛て先を座標としてＰＴの内容を参照することにより転送方向が決定される。ＰＴの内容は，負荷の調整が行われる度に更新されるが，更新の手間を軽減するために座標の下位ビットをマスクしてＰＴを引くことができるようになっている。即ち，ＰＥの受け持ち領域から「遠い」座標点に関するテーブルを「粗く」することにより，更新すべきエントリを大幅に減らすことができる。転送方向の競合などで生じるネットワーク中の渋滞を緩和するために，Ｃｈ０－３の出力には48×10ビットの出力バッファが設けられている。

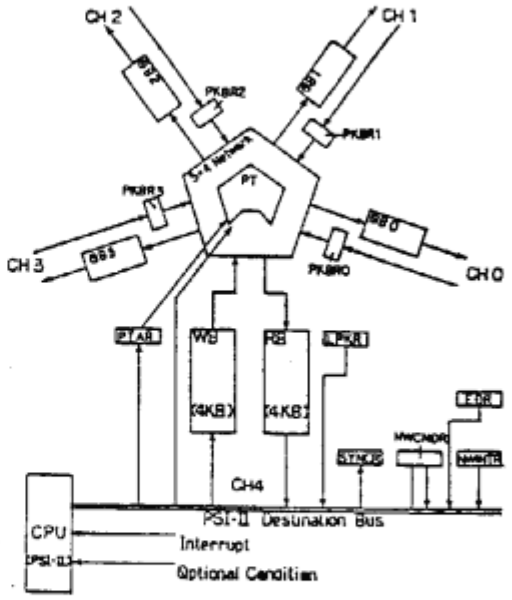


図６　接続用ハードウェアの構成

Ｃｈ４はＰＴを持たないため，ＣＰＵから送出されるパケットには転送方向を決めるために余分の２ビットが付加されている。またＣＰＵへ渡すパケットには受けとったチャネル番号を示す２ビットが加えられている。Ｃｈ４は，入出力とも４Ｋ×１２ビットのバッファを持ち，ＣＰＵから送出されるパケットはＷＢ(Write Buffer)中に全データが揃った時点でネットワークへの送出が開始され，ＣＰＵへのパケットもＲＢ(Read Buffer) 内に全データが揃ってからＣＰＵに対し到着が報告される。パケットの到着・送り出し及びネットワークの異常は，割込を用いてＣＰＵに報告される。

5　おわりに

本報告では並列推論マシンである，マルチＰＳＩ第２版に付いてネットワーク接続用ハードウェアを中心とし，その特徴とハードウェア構成について述べた。マルチＰＳＩ第２版の特徴としては，P^3を用いた動的な負荷分散方式があげられる。