7L−2

階層構造並列計算機のための多数の結合回路を持つプロセッサノードの設計

廣田　勝久　　高橋　義造
徳島大学工学部知能情報工学科
e-mail : hiro@n30.is.tokushima-u.ac.jp

1　はじめに

これまで、超並列計算機向きの相互結合網として、様々なものが研究されている。このうち多進木構造は、平均通信距離が短く、再帰的構造をもつためスケーラビリティに優れているなど、超並列向きのトポロジである。相互結合網として優れた特性を持つバイナリハイパーキューブの平均通信距離は、ｎをプロセッサ台数とした場合、

$d(HC(2)) \sim 0.5\log_2 n$

で表される。また、ｋ進木の平均通信距離は、

$d(k) \sim 2\log_k n$

で表わされる。これよりｋ進木の方が平均通信距離の短くなるｋを求めると、ｋ≧１６となり、１６進木以上では、平均通信距離について、バイナリハイパーキューブをしのぐ特性を持つことがわかる。

このように優れた特長を持った、図１のような多進木構造を持つ分散メモリ型並列計算機を構築するため、まず多数の接続ポートを持つプロセッサノード（ＰＮ）を設計した。　以下の節では、マシンを構成するプロセッサノードに付いて述べ、さらにルータ、Network Interface Unit（ＮＩＵ）の概要について説明する。

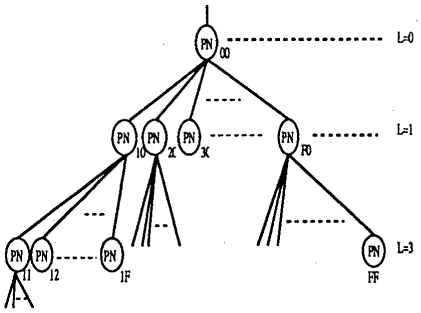

図１　１６進木構造を持つ階層構造並列計算機

2　プロセッサノード

ＰＮは、図２に示すように、ＭＰＵ、ルータ、そしてＭＰＵ−ルータ間のインターフェイスを受け持つＮＩＵから構成される。

Design of a processor node provided with many ports for hierarchical parallel computer.
Katsuhisa HIROTA and Yoshizo TAKAHASHI
Department of Information Science and
Intelligent Systems, University of Tokushima.

ＭＰＵには、６８Ｋをコアに持ち、周辺機能を１チップにまとめたＴＭＰ６８３０１（１２ＭＨｚ）を用い、メモリは２５６ＫＢのＳＲＡＭを実装している。

ノード間の通信は、専用のハードウェアであるルータとＮＩＵが受け持つ。これによって通信は高速におこなわれ、またＭＰＵの負荷も軽減される。

このような構造のＰＮを複数組み合わせることによって、１６進木構造の並列マシンを構築する。

木構造のマシンでは、リーフ（葉）の部分にあたるＰＮはルーティングを行わないのでルータが不要である。１６進木においては、リーフにあたるＰＮの割合が大きいので、ノード数に比較して、マシン全体におけるルータのハードウェア量が少なくて済む。

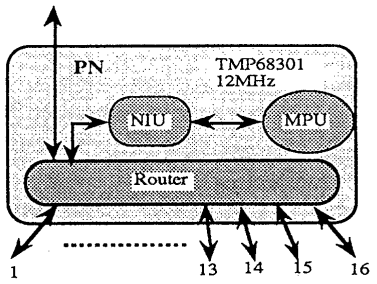

図２　プロセッサノード（ＰＮ）

3　ルータ

ルータは他のＰＮと接続する為のポートを多数持つ。１６進木構造のマシンのルータには、上位につながる１本、下位への１６本、及び自ノードに接続する１本の計１８の双方向のポートがつながることになる。

この様にルータには多数のノードがつながるので、高速な動作及び柔軟なフロー制御を実現することが要求される。

フロー制御の観点から、ＰＮ間の通信には、メッセージが通信回線を長時間占有する事のない様、パケット交換方式を採用し、バーチャルカットスルー（ＶＣＴ）[1][2]方式を用いて、ネットワークのスループットを上げ、レイテンシを短縮するなど、ネットワークの高性能化を図る。

ルータのハードウェアは、データバス幅８ビット幅の、１８ポート×１８ポートクロスバスイッチを中心に構成し、２０ＭＢ／ｓｅｃのデータ転送速度を実現する。

4 Network Interface Unit

ＮＩＵの構成を図３に示す。

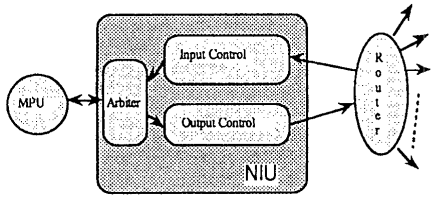

図３ Network Interface Unit

ＮＩＵはＭＰＵから受け取った送信メッセージを、パケットに変換してネットワークに送り出すＯｕｔｐｕｔ制御部と、ネットワークから受け取ったパケットを、元のメッセージに復元し、ＭＰＵに受け渡す働きをするＩｎｐｕｔ制御部、そしてこれらの各ユニットとＭＰＵ間のデータ転送の調停を行うアービタから構成される。各ユニットはＬＣＡを用いて実現する。

Ｏｕｔｐｕｔ制御部は、ＭＰＵからのコマンドによってＤＭＡを起動し、メッセージを一旦ＮＩＵ内のＦＩＦＯバッファに格納する。メッセージの最大長は４ＫＷとしている。これ以上の長さのメッセージを送るときは何回かに分けて送るようにする。

Ｏｕｔｐｕｔ制御部はメッセージを受け取ると、それをパケット単位に分割してネットワークに送り出す。パケットサイズは６７ワード固定長（ヘッダ３ワード、ボディ６４ワード）である。（１ワード＝１６ビット）

メッセージ送出の完了は、割り込みによってＭＰＵに伝えられる。

Ｉｎｐｕｔ制御部は１２８ＫＷの受信バッファを持っている。 ルータから自ノード宛に送られてきたパケットは、送信元ＰＮ、メッセージＩＤ別ごとに仕分けされ受信バッファ内に蓄えられて行く。

受信バッファのデータ構造を図４に示す。

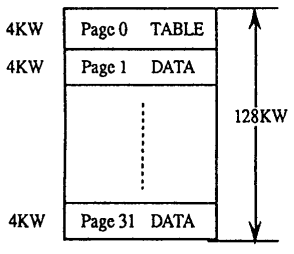

図４ 受信バッファの構造

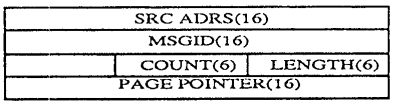

図５ テーブルの１要素

受信バッファは４ＫＷのページ単位で管理されている。このうち０ページは、現在保持しているメッセージの種類や状態などを記録しておくテーブル用の領域として用いる。テーブルの１要素を図４.３に示す。テーブル用のページを除いた残りの１ページから３１ページまでを、メッセージの格納用に用いる。このため、一度に保持出来るメッセージの個数は３１個以下に制限される。

メッセージを構成するパケットがすべて到着すると、Ｉｎｐｕｔ制御部はこれを検知し、ＭＰＵに割り込みをかけてこれを知らせる。メッセージを構成するパケットがすべてそろったかどうかの判断は、テーブルのＣＯＵＮＴとＬＥＮＧＴＨの値を見て行う。ＣＯＵＮＴは、今まで受け取ったパケットの個数で、ＬＥＮＧＴＨはメッセージを構成するパケットの個数である。この両者の値が等しくなれば、すべてのパケットが到着したと判断する。

並列計算機の通信網は、ＰＮ間の距離が分散システムなどに比べて短く、エラーレートも非常に低い。このため未到着のパケットがある場合には、タイムアウトなどによって、メッセージ全体を再送信するなどといった方法をとっても、さほどの効率低下にはつながらないと思われる。そのため、パケット再送などの、エラー訂正用の特別のハードウェア機構は設けていない。

メッセージが受信バッファ内に到着すると、ＭＰＵはＤＭＡを起動してこれを受信バッファから抜き取る。

ＭＰＵ側からは０ページが見える様になっており、到着メッセージのＩＤと、受信バッファ内での位置とサイズを知ることが出来る。ＭＰＵはこの情報を見てＤＭＡのパラメータをセットする。ＤＭＡ転送中は、パケットの受信は一時停止され、パケットはルータのバッファ内でブロックされる。

５．おわりに

現在プロトタイプを製作中であり、ＭＰＵ部分の製作を終え、ルータとＮＩＵの実装設計を行っている段階である。

参考文献

[1]W.Dally. : Network and Processor Architecture for Message-Driven Computers ,in VLSI and PARALLEL COMPUTATION,ed. R.Suaya,G.Birtwistlc. MORGAN KAUFMANN (1990)

[2]S.Konstantinidou,L.Snyder:Chaos router: architecture and performance. IN Proc. Int. Symp. on Comput. Arch., pp.212-221,(1991)