高並列推論エンジン実験環境ＰＩＥＥＥ
―自動負荷分散ネットワーク―

4P-5

山内　宗、　小池汎平、　野田　浩、　田中英彦
（東大　工学部）

1．はじめに

現在我々は、複数台の推論ユニットを、自動負荷分散ネットワーク［1］で結合した、ＰＩＥの並列処理向け実験環境ＰＩＥＥＥ（Parallel Inference Engine Experimental Environment）［2］［3］の製作を進めている。本稿では、高並列計算機の実現にあたって必要となる、各プロセッサへの効率的な負荷分散の機能を持つＰＩＥＥＥの自動負荷分散ネットワークについて述べる。

2．自動負荷分散ネットワーク

高並列計算機では、それぞれのプロセッサへの効率的な負荷の分配が重要である。ＰＩＥＥＥの自動負荷分散ネットワークは、タスクの転送と逆の向きに行先プロセッサの負荷情報を伝送し、負荷の量が最少の行先を選択して、そこにタスクを送りつけることにより、効率的な負荷の分配をするものである。現在製作している自動負荷分散ネットワークは、４入力４出力クロスバー・スイッチを基本とするスイッチング・ユニット（ＳＵ）を構成要素としたオメガ網であり、各ＳＵは、次段ＳＵから逆向きに伝送されてくる負荷情報を比較し、その中で最少の負荷情報を前段ＳＵに伝送する。そして、負荷分散をするときは、最少の負荷情報を伝えて来ている次段ＳＵへの経路を設定する。現在製作しているＳＵの設計方針を以下に列挙する。

① 負荷分散向けの転送モードと、行先を指定する通常の通信向けの転送モードをもつ。モードの切り替えは、ポート単位で行う。

② どちらのモードでもマルチキャスト（任意台数のプロセッサへの送信）ができる。

③ 多段結合網を対象とする。

④ ネットワーク内のクロック・スキューを考慮して、ＳＵ内は同期制御、ＳＵ間は非同期制御とする。

⑤ 回線交換方式をとる。データ転送は、同期でも非同期でも行える。

3．ＳＵのハードウェア構成

図１に、現在製作している自動負荷分散ネットワークの構成要素であるＳＵのブロック図を示す。各ブロックの機能を以下に列挙する。

① クロスバー・スイッチ

４入力４出力のクロスバー・スイッチであり、制御線dir で制御することにより双方向（半二重）の通信が可能である。伝送速度を速くするために、32ビット幅のデータ線でパラレルに伝送をする。また、未接続の出力側のデータ線ＯＰj からは、そこに接続することが可能な全ての行先プロセッサの中で最少の負荷情報（８ビット）が入力されている。そして、未接続のＯＰj からの負荷情報の中で最少の負荷情報を未接続のＩＰi から前段ＳＵへ伝送する。

② ルーティング部

ルーティング部は大きく分けてアドレス・コンバータとディストリビューション・セレクタの二つのユニ

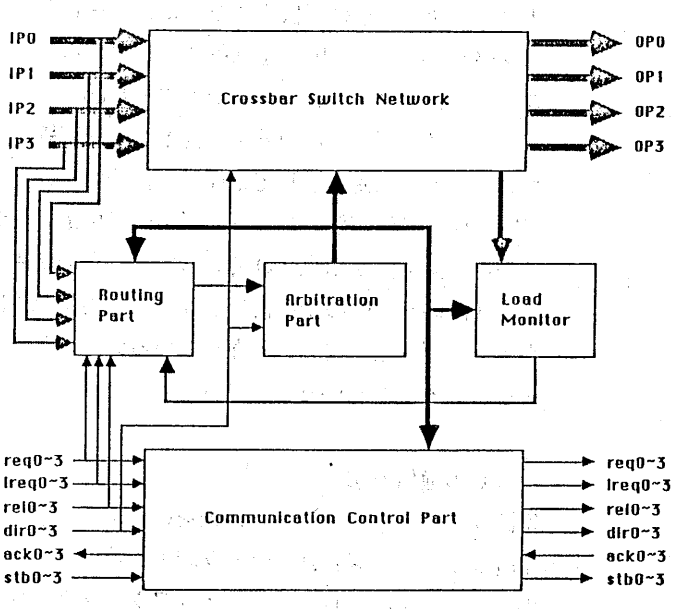

図１　ＳＵのブロック図

Parallel Inference Engine Experimental Environment PIEEE
-Automatic Load Balancing Network

Tsukasa YAMAUCHI, Hanpei KOIKE, Hiroshi NODA, Hidehiko TANAKA
University of Tokyo

ットで構成されている。アドレス・コンバータは、通常モードの際に、入力側のデータ線ＩＰｉの下位８ビットに多重化して伝送されて来る行先アドレスをデコードして、どのＯＰｊへ接続要求を出すかを決定する。ディストリビューション・セレクタは、負荷分散モードの際に、行先プロセッサの負荷情報を比較して選んだ、負荷の量が最少の行先への接続要求を出力する。そして、アドレス・コンバータあるいは、ディストリビューション・セレクタから出力された接続要求信号はアービトレーション部へ伝えられる。

③　アービトレーション部

ルーティング部からの接続要求が複数同時に生じることがあるので、アービトレーションが必要である。アービトレーション部では、優先順位を回転式で決定するアービターを使うことにより、接続要求のアービトレーションを公平に行っている。また、非同期に生じる接続要求に対応するために、ＳＵの内部クロック（20MHz ）で接続要求の同期をとっている。

④　負荷モニタ

未接続の出力側のデータ線ＯＰｊから伝送されて来る行先プロセッサの負荷情報を比較することにより、どのプロセッサの負荷が最少であるかを決定する。複数のプロセッサの負荷の量が同じである場合に、行先プロセッサを公平に選択するために、アービトレーション部と同様に優先順位を回転式にしている。

⑤　コミュニケーション・コントロール部

ＳＵには、データ線以外に、何本かの制御線が接続されている。制御線としてはreq （接続要求）、ack 、stb 、rel （接続解除）、lreq（モード指定）、dir （通信の向き）が用意されている。これらの制御線から、次段のＳＵへの制御信号を作り、入力側から出力側までの制御線の接続を管理しているのがコミュニケーション・コントロール部である。

４．ＳＵの通信方法

推論ユニット（ＩＵ）が、行先を指定する通常の通信の接続要求を出す場合は、データ線の下位８ビットに行先のアドレスを出力し、lreqをインアクティブ、req をアクティブにする。そして、行先のＩＵからack が返って来たら、req をインアクティブにして、stb （同期式）あるいはstb とack （非同期式）を使ってデータの伝送を開始する。また、負荷分散の通信の接続要求を出す場合は、lreqをアクティブ、req をアクティブにして、行先ＩＵからのack を待ち、それからreq をインアクティブにして、データの伝送を開始する。マルチキャストは上記の、接続要求を出してack を待つという動作を必要な接続数だけ繰り返すことによって実現される。そして最後に、接続を解除したい時は、rel をアクティブにする。

５．ＳＵの性能

現在製作しているＳＵは、ハードウェア量を減らし且つ動作速度を高速にするために、ＰＬＡ（Programmable Logic Array ）を多用した。ＳＵ各ブロック毎のハードウェア量を表１に示す。また、今回製作したＳＵの動作速度（アービトレーション部クロック周波数20MHz ）を表２に示す。一度接続が確定すると、ＳＵの内部クロックと無関係にデータ転送を行うので、転送レートはストローブ、アクノリッジ、データ転送のゲート遅延だけで決まる。

６．おわりに

ＰＩＥの並列処理向け実験環境ＰＩＥＥＥの、自動負荷分散ネットワーク及びその構成要素であるＳＵについて述べた。現在、製作中であり、なるべく、早期の稼働を目指したいと考えている。

表１　ＳＵ各ブロックのハードウェア量

	TTL	PLA
クロスバー・スイッチ	80	0
ルーティング部	0	8
アービトレーション部	0	14
負荷モニタ	19	6
コミニュケーション・コントロール部	5	20
合計	104	48

表２　ＳＵの動作速度

	最大遅延（ns）
行先指定の経路設定	160
負荷分散の経路設定	145
接続解除	145
ストローブ	15
アクノリッジ	25
負荷情報の更新	120
データ転送（順方向）	10
データ転送（逆方向）	10

<参考文献>

［１］坂井、小池、田中、元岡、”動的負荷分散を行う相互結合網の構成”、情処論、Vol.27,No.5,1986.

［２］小池、山内、田中、”高並列推論エンジン実験環境ＰＩＥＥＥの概要”、第33回情処全大、5B-5, 1986.

［３］小池、山内、野田、田中、”高並列推論エンジン実験環境ＰＩＥＥＥ　－全体構成－”、第34回情処全大4P-3, 1987.