PIEのモジュール間結合網

1R-3

坂井修一　小池汎平　田中英彦　元岡達

(　東京大学　工学部　)

1.はじめに

我々は現在，高並列推論エンジンPIEを開発中であるが，本マシンではモジュール間で大量のデータによる高頻度の通信が行われるため，転送性能の高い相互結合網の設計が必要となる。本稿では，PIEの相互結合網の実現方式とその設計に関して述べ，また，相互結合網を用いた動的負荷分散方式(2)の適用を検討する。

2.PIEにおける転送データと結合網の役割

PIEは2レベルより成る階層構成のシステムであり，下位レベル（レベル1）においてのみ構造データ（ Ground Instnce ）の共有を許している(1)。レベル1内で転送されるデータはゴールフレーム（GF），コマンド，構造データ，定義節，SMアドレスの5種類であり，各々の大きさと転送頻度はTable 1 に示すものとなる。Tuの値は，現在試作中のシステムで80μs程度，処理単位の変更などによる改良によって30μs程度になると予想される。また，レベル2では，GFおよびコマンドが転送される。

GF転送には，高スループット（1ポートあたり約10MB／s）と均等な負荷分散（後のタスク間の通信を考慮に入れたもの）が，コマンドと構造データの転送には早いレスポンス（数μs以内）が，定義節の転送にはマルチキャストの機能が要求される。

3.モジュール間結合網の構成

Table 1 に示した5種のデータのうち，GFと定義節は大きさの点が共通しており，他の3種のデータは大きさと要求（低遅延）が共通している。そこで，前二者・後三者にそれぞれ1つの結合網を割り当てる方式を採用することにした。本構成によって，各網の役割分担が明確になり，それぞれの通信機能が単純化され，転送の並列性も高まる。即ちレベル1システムは，GFと定義節を処理ユニットに分配する分配網（DN）と，AC間のコマンド転送・構造データ・SMアドレスの授受を行うコマンド・追加読み出し・アドレス転送網（CLAN）を持つことになる。同様にレベル2には，分配網（DN）とコマンド網（CN）がある。

3.1 DN（レベル1）

最もハードウェアコストの低いDNの実現法として，バスを用いる方式が考えられるが，GF転送に必要とされる容量の点から適用が困難である。一方，完全結合網やクロスバスイッチを用いれば高いスループットが得られるが，大きなハードウェア量が必要となる。

格子型網・超立方体網・CCC 網などの適用も考えられるが，これらは結合の局所性が大きく，問題をうまく分割して処理ユニットに割り当てないと，非局所的な通信が全体性能を低下させることになる。

レベル1のDNとしては，スイッチを付加して多ルート化したオメガ網を適用する。本網には結合の局所性がなく，したがってプロセッサ間距離の大きな転送が処理の隘路となる危険がない。また，ハードウェア量もO（N log N）と妥当であり，マルチパス化により信頼性も高い。

本網は，通常の通信機能の他に，GF分配のため自動的に負

Table 1. Transfer Data in Level 1 System

data		FROM	TO	size	freq.(/IU)*
GF		UP	MM	hundreds of bytes	≦ 1
		UP	IM		
		IM	MM		
struc-ture data	ΔG	UP	SM	10~20B	0.5~1
	LF com.	UP	SM	~10B	0.1~0.4
	LF data	SM	UP	10~50B	0.1~0.4
definition clause		UP	DSM	hundreds of bytes	<< 1
		DSM	DM		
SM address		SM	UP	~4B	< 0.3
		UP	SM		

* where T_u is the unit time
(T_u : execution time of 1 GF in 1 UP)
Size and frequency of communications are obtained by simulations.

荷分散を行う機能を持つ。即ち，本網は，図1に示されるスイッチング・ユニット（SU-2）を構成単位とするが，SU-2は，網の閉塞を回避しつつ，最少負荷量を持つ行先を自動的に選択して経路設定を行う機能が備わっており，これによって，(1)均等な負荷の分配，(2)ハードウェア的および時間的に低い負荷分散制御のオーバヘッド，(3)高い転送スループットが実現される。SU-2はマルチキャスト機能も持っており，ゲート数1740，入出力線数118 程度の4 ×4 の回線交換スイッチである(2)。

3.2 CLAN（レベル1）

CLANは，早いレスポンス（数μs）を要求される網であり，データの転送量はレベル1全体で最大約30MB／sとなる。また，処理ユニットの初期化の際などには，同時に全ユニットに情報を提供するのが望ましく，マルチキャストの機能が必要である。

低コスト・早い応答・マルチキャスト機能・SMが集中型であるなどの点から，レベル1のCLANには集中管理型個別要求論理バスを適用する（図2）。各処理ユニットは2本のrequest 線を持つが，これは追加読み出しとその他のデータ転送の間で優先度に差をつけるためである（前者が優先）。

アービタは動作速度と安定性の点から2階層のリングバッファ型のものを用いる。また，図中でEOD は転送データの終りを示す信号線である。転送速度向上のため，アービトレーションとデータ転送を重畳化する。

CLANに必要なバス幅は10B 程度である。Pollaczeck-Khintchineの公式を用いて解析を行ったところ，1回の追加読み出しに要する時間は，バスのコンフリクトによる待ちを含めて約4μs（およびSM内処理時間）となることが示された。

3.3 レベル2の結合網

レベル2のDNによって転送されるGFは，構造データを完全にコピーして持つものであり，レベル1のそれと比較して大きい。定義節およびコマンドは，レベル1のそれとほぼ同じ大きさであると考えられる。転送頻度は3者ともにレベル1のものより低いと予想される。

2種類の網（DN・CN）に関して，レベル1システム間でやりとりされるデータのトラヒックを見積り，その結果から実現方式の検討を行う予定である。DNとして3.1 で述べたマルチパス化した負荷分散適応型のオメガ網を，CNとして集中管理型個別要求論理バス（3.2）を適用することが考えられる。

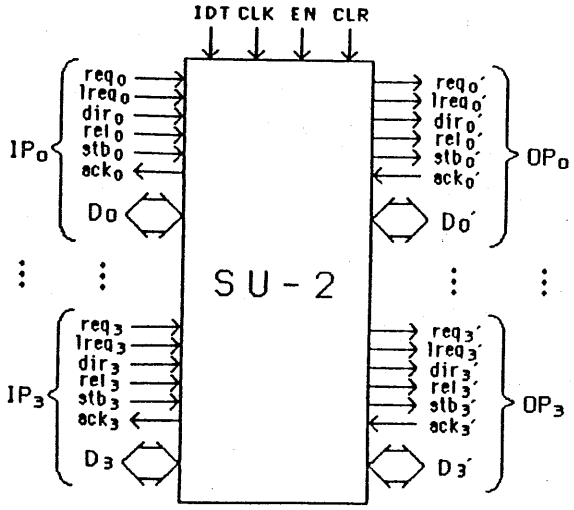


Fig. 1 Automatic Load Balancing SU (SU-2)

4.おわりに

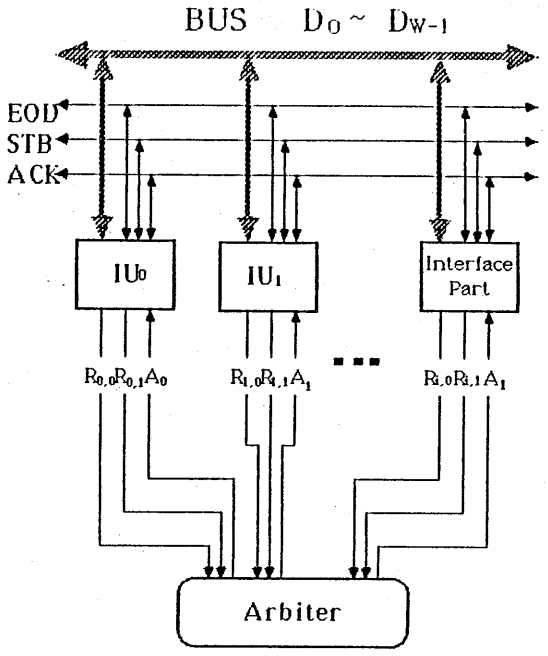
高並列推論エンジンＰＩＥの相互結合網の構成法に関して述べた。現在，3.1 で述べたSU-2（負荷分散適応型のスイッチング・ユニット）のLSI 化を検討中であり，また，3.2 で述べたバスの小規模のプロトタイプを試作中である。

今後の課題として，(1)レベル2のデータ・トラヒックの測定とレベル2の結合網の設計，(2)処理ユニット内の網インタフェース部の設計，(3)より大規模なシステムの相互結合網の設計などが挙げられる。

文献

(1)坂井，田中，元岡：“高並列推論エンジンＰＩＥにおける相互結合網の構成”，信学技報 EC84-46.

(2)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”，信学技報 EC85-24 .



STB : Strobe ACK : Acknowledge
EOD : End of Data $R_{k,0}$: Request from IU_k, Priority 0
$R_{k,1}$: Requst from IU_k , Priority 1
A_k : Acknowledge to IU_k

Fig. 2 CLAN Construction by Bus