1Q−8

PIE64の相互結合網の特性評価

高橋 栄一　小池 汎平　田中 英彦
{eiichi,koike,tanaka}@mtl.t.u-tokyo.ac.jp
東京大学 工学部

1 はじめに

並列推論マシンPIE64[1]は,記号処理プログラムの高並列・細粒度実行による高速処理を実現するために,特にリモートアクセスのレイテンシの小さい相互結合網[2]を採用している.

この相互結合網は,4×4のクロスバスイッチを基本構成単位とした多段網であり,回線交換方式を採用して双方向通信を実現している.また,シングルパスの多段網であるため閉塞状態の影響によるリモートアクセスのレイテンシの増大が懸念されるが,その影響を実用的な範囲に収めるために同一構成のネットワークを2系統用意した.本相互結合網は,さらに動的負荷分散を支援する機構を備えており,オーバヘッドなしに動的な負荷分散が実現できる.この機能は2系統のネットワークでそれぞれ独立に利用可能であり,2つを同時に使用することによってさらに効果的な動的負荷分散の実現が期待できる.

本稿では,PIE64の相互結合網の持つこうした特性を定性的に検討し,その有効性と妥当性について考察する.

2 相互結合網の構成と特徴

まず,PIE64の相互結合網の構成を概観し,その特徴について簡単に述べる.

PIE64の相互結合網は,図1に示すような3段の多段網である.したがって閉塞網であり,特定のPE間の通信パスは1通りしか存在しない.構成単位であるスイッチは,独自に開発したゲートアレイ1種4つから構成される32ビット幅,4×4のクロスバスイッチである.スイッチはバッファを持たず,全体として回線交換方式で通信が行なわれるため,データは電気的な遅延時間だけで双方向に目的のPEに伝達され[3],これによってレイテンシの小さいリモートデータアクセスが実現される.

しかし,この構成では通信量の増大にともなってリモートアクセスのレイテンシが急激に悪化してしまうので,実用的な範囲の通信量に対してリモートアクセスのレイテンシを低く保つために,こうしたネットワークを2系統用意した.

また,本ネットワークは[4]で提案されている動的負荷分散支援機能を採用しており,わずかなハードウェアの追加で効果的な動的負荷分散の達成を可能としている.

以下に,作製したPIE64の相互結合網[5]の特徴をまとめる.

- 64入力 × 64出力,3段の多段網
- 各通信パスは38ビット — 双方向通信が可能なデータ線が32ビット,通信制御線が6本

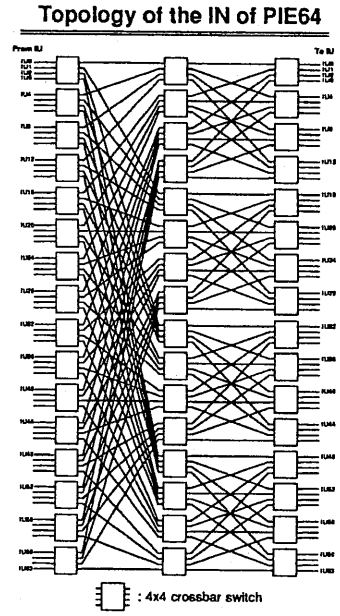

図 1: 相互結合網の構成 (1系統分)

- 同一構造の独立した2系統のネットワーク
- 10MHzの単一システムクロックに同期して動作
- 動的負荷分散支援機能
- バンド幅は1ポート当たり40 $MBytes/s$,1系統で2.56 $GBytes/s$, 全体(2系統)で5.12 $GBytes/s$
- データや制御信号の伝達遅延は約60 ns (1クロック以内)
- 負荷情報の更新に約60 ns (1クロック以内)
- データ転送方向の反転に約60 ns (1クロック以内)

3 リモートアクセスのレイテンシ

一般に相互結合網の設計においては,リモートアクセスのレイテンシよりもプロセッサ間通信のスループットの方が重視され,また,スループットを向上させることによりレイテンシの影響を抑える試みがなされる.確かに,ある程度粒度の大きいプロセスを並列実行し,プロセッサ間通信時の転送データ長は大きいが頻度は低いと言う場合には,リモートアクセスの待ち時間に他のコンテクストに対する処理を行なうことにより,実質上待ち時間を見えなくすることができる.しかし,各プロセスの粒度が小さい場合には,それらのプロセスが頻繁に「粒度の小さい」リモートアクセスを行ない,結果としてレイテン

Characteristics of the interconnection network of PIE64
Eiichi TAKAHASHI, Hanpei KOIKE, Hidehiko TANAKA
The Faculty of Engineering, The University of Tokyo

シの影響を隠すことができなくなってしまう.

また,1回のリモートアクセス中にReadとWriteが同時に発生するような場合も考慮する必要がある.リモートアクセスがWriteだけであれば, Wormhole Routingなどでも十分小さいレイテンシが達成できるが,リモートデータのRead時に実際に生じるトラフィックは,

Write(アドレスを送る) ⟹Read(リモートデータを得る)

であり,Read-Modify-Writeであれば,

Write(アドレスを送る) ⟹Read(リモートデータを得る)
⟹Write(書き戻す)

となる.こうしたReadおよびWriteの複合アクセスのレイテンシを小さくするには,回線交換方式で双方向通信を行なうのがハードウェア的に見て最も容易な実現方法である.

リモートアクセスのレイテンシを増大させる要因には

- 相互結合網のスイッチ内における通信要求の衝突
- 目的のPEの入口でのアクセス要求の衝突
- 目的のPE内での待ち時間
- 相互結合網内の電気的な信号伝搬時間
- 通信経路設定/解除のオーバヘッド
- 要求元のPE内でのリモートアクセス要求の生成時間
- 転送データ長(1セッションの所要時間)
- データ転送レート

などが挙げられる.これらに対してPIE64では,まず相互結合網ハードウェアレベルで次のような対策を立てた

- 実装構成を工夫して相互結合網内の伝搬遅延時間を1クロック以内とし,PE間での同期転送を可能にした.
- 経路設定/解除に要するクロック数を小さくし,オーバヘッドを低減した.

ネットワークの多重化,および,動的負荷分散支援機能の効果については次節以降で述べる.この他に,PE側ではリモートアクセス処理専用のハードウェア[6]を用意し,アクセス要求の生成,および,受付処理を高速化している.他の項目は,負荷の分割/分散の戦略によって解決されるものであり,現在検討を行なっている.

4 動的負荷分散支援機能

まず, PIE64が採用している動的負荷分散支援機構について概説する.

各PEは通信に使用していないパスを用いて自己の負荷量をネットワークの出力側から入力する.ネットワーク内の各スイッチは,通信に使用していないパスから送られてくる負荷量を比較し最小値を前段に送るとともに,最小負荷を送ってきたパス(スイッチのポート)を記憶する.こうして通信に使用していないネットワークの入力ポートからは,その時点の全PE中の最小負荷が観測され,また,ネットワークはその最小負荷を持つPEへの通信経路を即座に設定することが可能となる.

この機能により,動的負荷分散のための負荷情報の管理が不要になり,要求時に即座に通信経路を設定できるため,通常の通信と同じコストで経路設定やデータ転送が実現できる.また,最小負荷値が常時観測できるので,負荷分散すべきかどうかの判断が瞬時にできる.したがって,負荷情報の交換や最小負荷PEの確認などのための通信が一切不要であり,トラフィックの低減にも貢献している.

この機構は, PIE64では約6000ゲートのゲートアレイのうちの1000ゲート弱で実現されている.全体に占める割合としては大きいかも知れないが,スイッチ素子としての規模が小さいためであり,絶対量としてはそう多くはない.

5 ネットワークの二重化

トラフィック増大時におけるリモートアクセスのレイテンシの悪化を改善するために, PIE64ではネットワークを二重化し,スイッチ内,および,PEの入口でのアクセス要求衝突の軽減を図った.

この方法はハードウェア的には最もコストのかかる方法であるが,実現が可能であるならば最も有効な方法の一つである.これによりトラフィックの多いところでのレイテンシが改善されるだけでなく,トラフィックの分散によって通信要求の衝突する確率が減少し,通信トラフィックの全域に渡ってレイテンシが改善される.

また,前節で述べた負荷分散支援機能は,通信に使用されていないパスを用いて負荷情報を伝達するので,通信中のPEからの負荷情報は考慮できず,相互結合網側が持っている最小負荷情報は必然的にある程度の誤差を含んでいる.しかし,ネットワークが2系統あれば,相対的に負荷情報の伝達に利用できるパスが増え,最小負荷情報の精度向上が期待できる.別の扱い方として,負荷情報を2次元の値として表現する,つまり,各PEが2系統のネットワークに対して異なった意味を持つ負荷量を提供することにより,より詳細で的確な負荷分散の判断が可能になる.

6 おわりに

本稿では, PIE64の相互結合網が持つ特性についての考察を行なった.本相互結合網については,さらにシミュレーションによる定量的な検討や,既に作製済みのハードウェアを用いた測定に基づく考察を行なう必要があり,ここでの評価は最終的なものではない.また,実際には要素プロセッサと合わせた評価も必須であると考えており,要素プロセッサの完成を待って行なう予定である.

参考文献

[1] 小池 汎平, 田中 英彦: "並列推論エンジンPIE64", 並列コンピュータアーキテクチャ, bit臨時増刊, Vol.21, No.4, 1989, pp. 488-497.

[2] Koike H., Takahashi E., Yamauchi T. and Tanaka H.: *The High Performance Interconnection Network of Parallel Inference Machine PIE64,* Computer Architecture Symposium IPS Japan, 1988.

[3] 高橋, 小池, 田中, "PIE64の相互結合網の電気的特性の評価", 第41回情報処理学会 全国大会, Sept. 1990.

[4] 坂井, 小池, 田中, 元岡, "動的負荷分散を行なう相互結合網の構成", 情処論, Vol.27, No.5, 1986.

[5] 高橋, 小池, 田中, "並列推論マシンPIE64の相互結合網の作製および評価", 情処論, Vol.32, No.7, 1991.

[6] 清水, 島田, 小池, 田中, "PIE64のネットワーク・インターフェース・プロセッサLSIの詳細", 情報処理学会, 計算機アーキテクチャ研究会 87-5, March, 1991.