

超並列計算機のネットワークの実現可能性と性能評価

斉藤哲也 森本貴之 位守弘充 朴泰祐 中村宏 中澤喜三郎

筑波大学 電子・情報工学系

Email: saito@arch3.is.tsukuba.ac.jp

分散メモリ型超並列計算機のための数種類の代表的なプロセッサ間結合ネットワークについて、それぞれの性能と必要ハードウェア量の観点から、定量的に比較を行なった。性能に関しては、主に大規模科学技術計算の並列処理を行なうときに比較的よく現れる、数種類のプリミティブな転送パターンについての転送レートを求めた。ハードウェア量に関しては、単純にスイッチ数を数えるのではなく、それぞれのネットワークを実装するのに必要なLSI数を取りあげ、ネットワークの転送データ幅をもとに、LSIのピン数の制約条件から必要LSI数を求めた。プロセッサ数については、実現可能な台数としてここでは一応1024台を仮定した。これらのデータにより、現実的な実装を考慮した場合、プロセッサ間結合ネットワークのコスト/パフォーマンス比較の一視点を提示した。

Quantitative Cost-Performance Evaluation of Networks for Massively Parallel Processing

Tetsuya SAITO Takayuki MORIMOTO Hiromitsu IMORI
Taisuke BOKU Hiroshi NAKAMURA Kisaburo NAKAZAWA

Institute of Information Sciences and Electronics
University of Tsukuba

In this research, we evaluate the cost-performance of several typical networks for massively parallel processing system from the view point of hardware amount for construction. As to performance of network, we evaluate the data transfer rates for several primitive data transfer patterns which often appear in solving scientific problems. As to hardware amount, we count the number of required LSI's instead of the number of switches. It is determined from the constraints of the number of signal pins per chip and the data bus width between switches. The assumed number of processors is 1024, which is a reasonable number to build a machine with the current technology. According to the evaluation result, we can discuss on the cost-performance comparison of these networks.

1. はじめに

近年、高性能なマイクロプロセッサ数千台を相互結合して飛躍的な処理速度の向上をねらう、超並列計算機に対する期待が高まっており、研究も活発に行なわれている。

このような超並列計算機の性能は、個々のノードプロセッサの能力及びノード間データ転送能力により決定される。特に後者は重要で、いくらノードプロセッサに性能のよいものを使っても、ネットワークの能力が低ければ、計算機全体の能力は出ず、膨大な資源の無駄使いとなる。さらに、ネットワークの性能がよくても、あまりにもコストがかかるのであれば、そのネットワークの採用は現実的に不可能である。

今後、このような超並列計算機が次々と作られ商品化されていくと思われるが、その際、コストとパフォーマンスに大きく影響するという点で、どのネットワークを採用したらいいか、ということが大きな問題となると思われる。

そこで、本報告では、分散メモリ型超並列計算機のための数種類の代表的なプロセッサ間結合ネットワークについて、それぞれの性能と必要ハードウェア量の観点から、定量的に比較を行なった。ハードウェア量に関しては、必要とされるスイッチ数に対する評価は従来より行なわれているが、本研究ではネットワークの実装に必要なLSI数という観点からこれを評価した。なぜなら、現実的な実装を考慮した場合、スイッチはLSIの中に実現されることになり、必要となるLSI個数は、LSIのスイッチ用回路の数よりもLSIの入出力ピン数による制約によって決まってしまうからである。そこでピンネックを元に、各ネットワークを実装するために必要となるLSI数を求め、ハードウェア量とした。

評価の対象としたネットワークは、メッシュトラスネットワーク、ハイバキューブネットワーク、ハイバ・クロスバ・ネットワーク [1]、ADENART計算機で採用されているハイバクロスネットワークである [2]。以下、これらをそれぞれ、MT、HC、HXB、ADENAと呼ぶことにする。また本研究では、MTとHXBについては2次元構造と3次元構造を考えた。

これらのネットワークに対して、主に大規模科学技術計算の並列処理を行なうときに比較的よく現れる、数種類のプリミティブな転送パターンについての転送レートを求めた。ネットワーク性能は、転送所要時間がデータ量に比例する転送レートと、データ量に無関係なオーバーヘッドによって評価される

が、ここでは転送レートのみを性能を表す尺度として用いることにした。評価に用いた基本転送としては、隣接ノード間データ転送、斜め転送、桂馬転送、reduction、1対n broadcast、n対n broadcast、FFT (バタフライ転送)、2次元arrayの転置、3次元arrayの転置である。

これらのデータを元に、現実的な実装を考慮した場合の各ネットワークのコスト/パフォーマンスを比較検討する。

2. 評価ネットワーク及び評価条件

2.1 評価ネットワーク

ここでは、本研究で仮定した計算機の構成について述べる。

プロセッサは全部で1024台あり、1ノードに1プロセッサの構成とする。メモリは各ノードごとにローカルに置かれ、各ノードとネットワークを接続するためにNI (Network Interface) を設ける (図2.1)。

次に、評価するネットワークについては、表2.1のものを考える。

ネットワークの転送データ幅については、1 Byte、2 Byte、4 Byteの全二重転送路を仮定し、それぞれの場合について、上記ネットワークを実装するのに必要なLSI数を求めた。パリティビットについては1 Byte、つまり8 bitごとに1 bitついているものとする。

ノード間データ転送としては、sourceノードからdestinationノードへ、転送経路の途中にあるノードのメモリにデータをバッファリングしながら1段1段転送を繰り返す、ストアアンドフォワード方式 (以下、s/f方式と呼ぶ) と、sourceとdestinationの間で他のノードを介さずに直接転送できるウォームホール方式 (以下、w/h方式と呼ぶ) を考えた。両者とも、転送ピッチは40[ns]を仮定した。

図2.2、図2.4において、EX (エクスチェンジャ) と呼ばれる回路は、直接結合されていないノード同士を結合するための回路で、w/h方式を支援するためのルータ・スイッチである。EXも小規模なクロスバスイッチである。w/h方式では、sourceとdestinationの間の転送路が一度確保されれば、その転送が終了するまでその転送路を占有できるものとした。また、HCとADENAについては、既存のマシン [2] [3] を参考にs/f方式のみを考える。

2.2 評価条件

本研究では、大規模科学技術計算によく現れる基

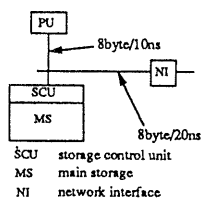


図2.1 ノードの構成

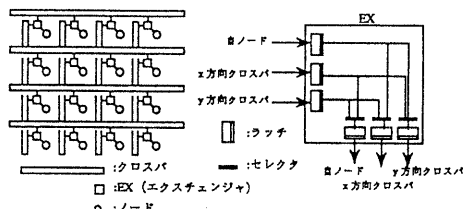


図2.2 w/h方式HXB (2次元)

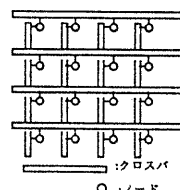


図2.3 s/f方式HXB (2次元)

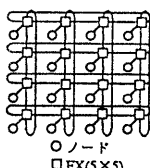


図2.4 w/h方式MT (2次元)

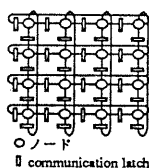


図2.5 s/f方式MT (2次元)

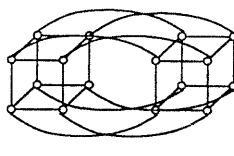


図2.6 s/f方式HC

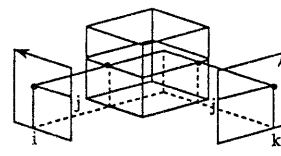


図2.7 s/f方式ADENA

トポロジ	次元数	ルーティング方式	ノード構成	図番号
HXB	2次元	w/h方式	32×32	(図2.2)
		s/f方式	32×32	(図2.3)
	3次元	w/h方式	8×8×16	
		s/f方式	8×8×16	
MT	2次元	w/h方式	32×32	(図2.4)
		s/f方式	32×32	(図2.5)
	3次元	w/h方式	8×8×16	
		s/f方式	8×8×16	
HC	10次元	s/f方式		(図2.6)
ADENA	2次元	s/f方式	32×32	(図2.7)

表2.1 評価ネットワーク

本転送を問題としているので、各ノードに割り当てられる仕事量が大きく、かつ、他ノードへのデータ転送量も大きい場合を想定した。データ転送量が多いので、ノードが転送の準備にかかる時間や、回線確保のためにかかる時間などは、データ転送そのものに要する時間に比べ無視できるほど小さいと考えられる。そこで本研究ではネットワークスループ

ットのみを問題とした。

また、各ノードへ割り当てられる仕事の量はすべて等しく、各ノードが結合ネットワークを通して他ノードへ転送しなければならないデータはすべて同一のデータ長であると考えた。

さらに、議論を簡単にするために、処理の進め方としては、すべてのノードが完全に同期して、計算フェーズ、転送フェーズがきれいに分離しているようなモデルを考えた。

評価に用いた基本転送パターンは以下の通りである。

- ・隣接転送
- ・斜め転送
- ・桂馬転送 (将棋の桂馬の位置関係の転送)
- ・reduction
- ・1対n broadcast
- ・n対n broadcast (complete exchange)
- ・FFT (バタフライ転送)
- ・2次元arrayの転置
- ・3次元arrayの転置

マッピング及びルーティングは、各ネットワークごとに、転送距離及び経路上での衝突が最小となるようにアルゴリズムを最適化した。

3. ハードウェア量

ここでは、2節で述べた各ネットワークを実装すると仮定したときに必要となるハードウェア量について考える。ハードウェア量は、より現実的な実装を考慮し、単にスイッチ数で数えるのではなく、

ネットワークを構成するために必要となるLSI数で求めた。

図3.1から図3.6に2節で示した各ネットワークを構成するために必要となる構成要素を示す。なお、HXBとMTの3次元構造の構成要素はそれぞれ、2次元構造と同様であるから、ここでは図を省略する。

これらの構成を実現するのに必要なLSI数を求める際の方針を示す。まず、1つのLSIの、信号線に対するpin数は350とする。実際には、この他に制御線及びかなりの本数の電源・アース線が必要となるので、現在の技術ではこのあたりが限界であると思われる。その上で、各ネットワークについて、転送データ幅が1 Byte、2 Byte、4 Byteのときについて各LSIに入る信号線の数を数える。1つのLSIに入る信号線の数が350を越えた場合は、LSIをbit-slice的に横に並べ、信号線数が350以内になるようにする。

信号線数が超過しているLSIのbit-sliceの方針として、信号線は1 Byte、2 Byte、4 Byte単位でしか切らない。これは、パリティビットがByte単位についているためである。しかし、bit-sliceを行なって信号線を1 Byte (9 bit) 切りにしても、そのLSIに入る信号線の総数が350を越えてしまう場合には、さらにLSIを分割し、パリティについてはその1 Byteのデータを受け持つLSIのグループ単位で処理する。

以上の方針に従ってLSI数を求めればよいが、図3.1から図3.6に示したそれぞれの場合について、簡単に

説明する。

・w/h方式HXBの場合 (図3.1)

w/h方式HXBでは、cross-barを実現するためのLSIとEXを実現するためのLSIの数を数えればよい。しかし、ノード内のSCUをLSIで実装することも考え、SCUとEXを1チップ上にまとめ信号線を有効に使うことを考える。そのとき、w/h方式HXBのハードウェア量はcross-barを実現するためのLSIのみになる。

・s/f方式HXBの場合 (図3.2)

s/f方式HXBの場合は、cross-barを実現するためのLSIの数を数えればよいので、必要なLSI数はw/h方式HXBと同じになる。

・w/h方式MTの場合 (図3.3)

w/h方式MTの場合は、meshの格子点上にあるEXを構成するLSI数を求める。この場合EXは、2次元構造の場合は 5×5 の、3次元構造の場合は 7×7 のcross-barである。

・s/f方式MTの場合 (図3.4)

s/f方式MTでは、転送路上にあるcommunication latchを実現するのに必要なLSI数を求めればよい。

・HCの場合 (図3.5)

HCでは、各ノードが10個のノードと双方向につながっており、ノードごとに、10入力のマルチプレクサと10出力のエクスパンダが必要となる。

・ADENAの場合 (図3.6)

ADENAは、 32×32 のクロスバスイッチが32枚積み

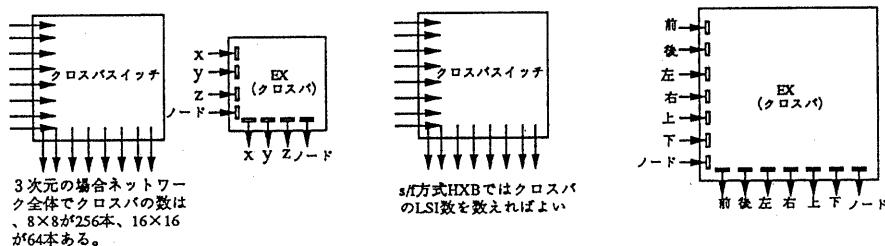


図3.1 w/h方式HXBの場合 (3次元)

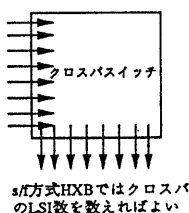


図3.2 s/f方式HXBの場合

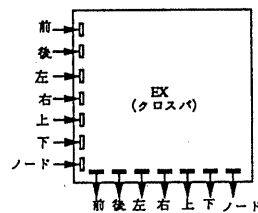


図3.3 w/h方式MTの場合 (3次元)

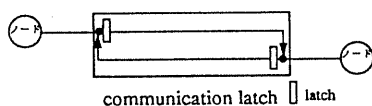


図3.4 s/f方式MTの場合

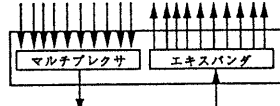


図3.5 HCの場合

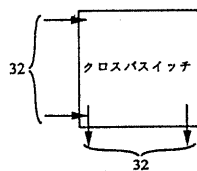


図3.6 ADENAの場合

重なった構造をしている。

以下の表3.1に各ネットワークの実装に必要なLSI数をまとめて示す。

トポロジ	次元数	ルーティング方式	転送データ幅 [Byte]		
			1	2	4
HXB	2次元	w/h方式	192	384	768
		s/f方式	192	384	768
	3次元	w/h方式	320	384	768
		s/f方式	320	384	768
MT	2次元	w/h方式	1024	1024	2048
		s/f方式	2048	2048	2048
	3次元	w/h方式	1024	1024	2048
		s/f方式	3072	3072	3072
HC	10次元	s/f方式	1024	2048	4096
ADENA	2次元	s/f方式	96	192	384

表3.1 各ネットワークの実装に必要なLSI数

4. 転送能力

ここでは、2節で示した様々な並列処理によく現れる基本転送を、各ネットワーク上で行なったときの転送レートを求めて、ネットワークの転送能力について検討する。2節でも述べたように、データ転送量が大きい場合を想定しているので、ネットワークスループットのみを問題とした。また、各ノードが他ノードへ転送しなければならないデータ長はすべて同一であるとした。

以下、それぞれの転送パターンについて、各ネットワークごとの転送レートを考える。ここでの転送レートとは、1つのノードあたり $N[\text{byte}]$ の送出データをもって、全ノードの転送が $T[\text{sec}]$ で完了した

とき、そのネットワークの転送レートは $T/N[\text{sec/byte}]$ であるとした。つまり T/N の小さいネットワークが優秀であることになる。なお、バスのスループット tp は、ネットワークの転送ピッチが $40[\text{ns}]$ であり、転送データ幅は $1[\text{byte}]$ 、 $2[\text{byte}]$ 、 $4[\text{byte}]$ を仮定したので、それぞれ、 $tp=1/40[\text{byte/ns}]$ 、 $1/20[\text{byte/ns}]$ 、 $1/10[\text{byte/ns}]$ となる。

4.1 隣接転送

問題の格子空間の計算機へのマッピングは、HXB、MT、ADENAについては、ノードが構成する格子に素直にマッピングした。HCは、グレイコードで表したMTを想定しマッピングした。これらのマッピングは斜め転送、桂馬転送、arrayの転置のときも同じである。

ここで考えた隣接転送は1024台すべてのノードがすべて同じ方向に同時に転送する場合である。従って、データ転送路上での衝突は生じない。

隣接転送は、w/h方式でもs/f方式でも同様に転送でき、隣接転送にかかる転送時間 T は、HXB、MT、HCについては、 $T=1/tp \times N$ 、ADENAについては、 $T=1/tp \times N \times 2$ となる。それぞれの転送レートは前者が $1/tp$ 、後者が $1/tp \times 2$ となる。

4.2 斜め転送

斜め転送も1024台すべてのノードがすべて同じ方向に同時に転送する場合である。斜め転送は、s/f方式ネットワーク、ADENAでは、2段の転送が必要である。w/h方式ではEXが有効に働き、衝突なしに直接転送できる。

4.3 桂馬転送

桂馬転送も同様に、1024台すべてのノードがすべて同じ方向に同時に転送する場合である。

w/h方式HXBでは衝突なしに転送ができる。s/f方式HXBでは、先に横方向に隣接転送を行ない、次に2つはなれたノードの転送を行なう。w/h方式MTでは、隣接転送と斜め転送を行なうことで桂馬転送を行なう。s/f方式MTは、隣接転送を3回繰り返して桂馬転送を行なう。ADENAでは、2段の転送で桂馬転送が完了する。

4.4 1対n broadcast

2次元w/h方式HXBは、図4.1のようにしてbroadcastを行なえばよい。3次元のときも同じ考え方で行なう。

このとき、転送レート T/N は $T/N=1/tp \times \log_2 1024$ と

なる。s/f方式HXB、w/h方式MT、HCもまったく同じになる。ADENAでは [4] の方法を用いることで全く同様の転送レートになる。s/f方式MTでは、途中にあるいくつかのノードを経由しなければならないので、その分遅くなり、 $T/N=1/tp \times (\text{ネットワークの直径})$ となる。

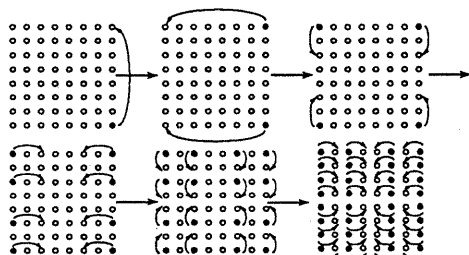


図4.1 2次元w/h方式HXBでの1対n broadcast (8x8の場合)

4.5 n対n broadcast (complete exchange)

すべてのノードが自分以外の他のすべてのノードに同時にbroadcastするので、この基本転送を完了するための一番簡単な方法は、1024個のノードを1回ずつ通るような一筆書きを行ない、リングネットワークを構成し、全ノードがリング上で一斉にデータ転送を行なって、データを1023ノード分循環させることである。ADENA以外のネットワークではこれは簡単である。また、ADENAの場合は例えば [5] のような方法により作ることができる。このようにリング上をデータを循環させれば、n対n broadcastの転送レートはどのネットワークでも同じになり、 $T/N=1/tp \times 1023$ で求められる。

4.6 reduction

ここでいうreductionとは、例えば全ノードのデータの総和を求めるような操作で、基本的には、データ転送とそれに対する何らかの演算の組み合わせを順次行なうものである。

このreductionを行なうには1対n broadcastの逆のことを行なえばよい。ただ、データを集める際に自分のもっているデータと転送されてきたデータで演算を行なうので、その演算時間も考慮しないといけない。この演算時間は転送データ長N[Byte]に比例すると考えられるのでkを比例定数としてkNとする。

従って、s/f方式MTでは転送レートは、 $T/N=1/tp \times (\text{ネットワークの直径}) + k \times (\text{ネットワークの直径})$ となる。その他のネットワークでは $T/N=1/tp \times \log_2 1024 + k \times \log_2 1024$ となる。

4.7 FFT (バタフライ転送)

ここで考えたFFTのモデルは、ネットワークによって異なる。まず、ノード構成が $8 \times 8 \times 16$ である3次元HXB、3次元MTでは、x方向3段のFFT、y方向3段のFFT、z方向4段のFFTを独立に行なう。なお、HCもこの3次元モデルで評価した。そして1ノードに割り当てられる標本点の数を $m \times m \times m$ とした。つまり、問題の空間の大きさは $8m \times 8m \times 16m$ である。

FFTでのバタフライ転送では、データを転送するたびにノードごとに演算を行なう必要がある。その時間は1回のバタフライ転送で1つのノードが送出する転送データ長Nに比例すると考えられるので、 $k \times N$ とする。また、バタフライ転送のあと、各ノードごとに残りのFFT処理を行なう。これにかかる時間をhとする。バタフライ転送での衝突、演算時間kNとhを考慮しながらFFT処理に必要な時間を考えた。

2次元HXB、2次元MT、ADENAでは、x方向5段のFFT、y方向5段のFFTを独立に行なうときの転送時間を求めた。

4.8 2次元array転置

要素数が $P \times P$ の2次元arrayを図4.2のように、各ノードに割り当てる。1ノードあたりの転送データ長をN[Byte]とする。したがって、ある1ノードから、ある1ノードへ転送しなければならないデータ長は $N/(1024-1)$ である。2次元arrayの転置を行なうには、各ノードは自分以外の1023個のすべてのノードにデータを転送しなければならない。HXB、MT、HCでは、このこととネットワークの対称性を活かしてデータ転送を行なう。つまり、ある1つのノード

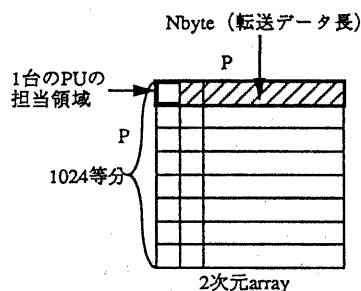


図4.2 2次元arrayのマッピング

4.9 3次元array転置

HC : $T/N = 1/t_p \times 1/32 \times 32/2 \times \log_2 1024$

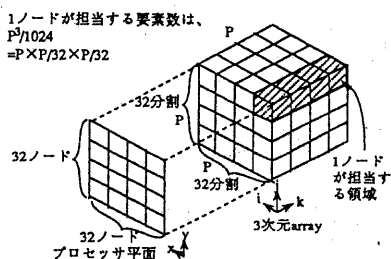


図4.3 3次元arrayのマッピング

5. コスト/パフォーマンス

今回はノード数として1024を仮定したが、将来的にはより多くのノードをもつ並列計算機が実現できるようになるであろう。その際、ADENAでは 32×32 以上のクロスバスイッチを実装しなければならない。一方w/h方式HXBでは、次元数を増やすことができるのでノード数が増えても対応することができる。LSI数はw/h方式HXBの方が比較的多く必要であるが、このような拡張性を考えた場合、w/h方式HXBの方が優位であると考えることができるだろう。

6. おわりに

その結果、ADENA、HXBのコスト/パフォーマンスが良いということがわかった。特にHXBは、MT等に比べて物量が多いといわれるが、それはスイッチ数で比べた場合の話であり、本研究のように現実のハードウェア量に近いLSI数で比べると、全く逆の結果が出た。

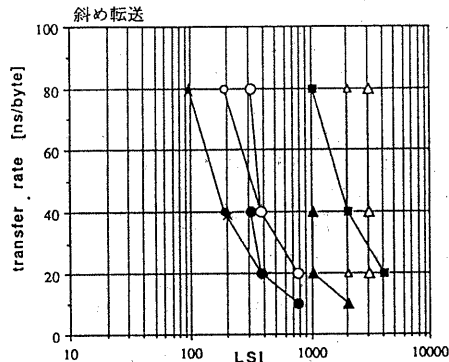


図5.1 斜め転送の場合
(隣接転送、桂馬転送、 n 対 n broadcastも
ほぼ同様の傾向を示す)

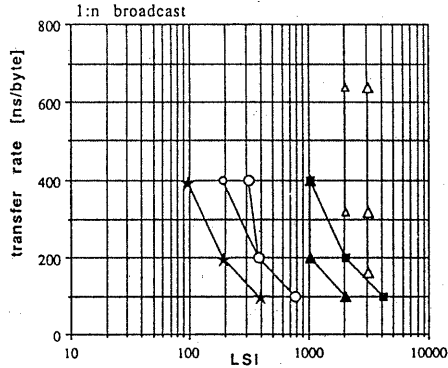


図5.2 1:n broadcastの場合
(reductionもほぼ同様の傾向を示す)

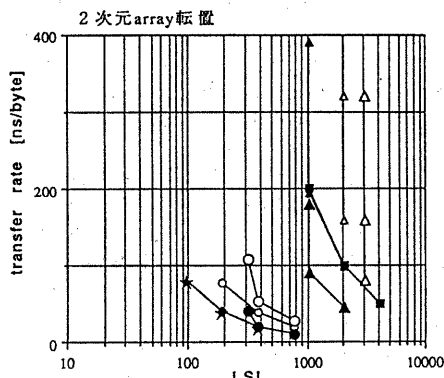


図5.3 2次元array転置の場合

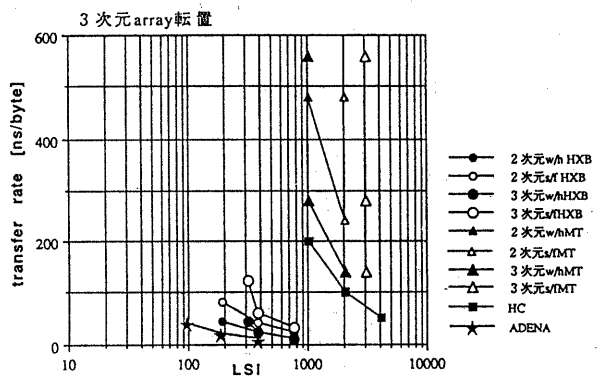


図5.4 3次元array転置の場合
(バタフライ転送もほぼ同様の傾向を示す)

今後の課題としては、より詳細に現実イメージに近いモデルで、多くのケースの評価を行なう必要がある。ハードウェア量としては、LSI数を信号線のピンネックだけから求めるのではなく、LSIの内部まで詳しくモデルを考慮して、評価するべきであろう。またネットワーク性能の評価については、本研究で行なった基本転送以外にも、連立一次方程式や行列の積、ランダム転送などの、並列処理によく現れる様々な基本操作に対する評価を行なう必要がある。また転送に関する性能については、各種オーバーヘッドも考慮するべきであろう。さらに、評価するネットワークとしてCM-5に採用されているといわれているFat-Treeなどの興味深いネットワークについても検討を行なっていきたい。

参考文献

- [1] 中越順二, 田中輝雄, 濱中直樹, 面田耕一郎: 並列計

算機H2Pの要素プロセッサ間非同期データ転送方式, 情報処理学会第38回全国大会, 6T-7, 1989

- [2] 岡林一郎, 狩野信吾, 貴志哲司, 中倉康浩, 瀬川礼二, 渡辺征男, 材木幸治, 金子克幸: 並列計算機ADENAにおけるネットワークの構成とVLSIによる実現, 信学技報ICD89-152, 電子情報通信学会, 1989

- [3] W. Daniel Hillis: The Connection Machine, The MIT Press, 1985

- [4] 西川順二, 野木達夫: 2進木構造を高速処理可能な並列計算機アーキテクチャについて, 情報処理学会第36回全国大会, 2C-4, 1988

- [5] 西川順二: 並列計算機ADENARTにおける各種相互結合網の実現, 並列処理シンポジウムJSP92, pp.227-234, 1992