SSS 型 PBSF(Piled Banyan Switching Fabrics) チップ

笹原 正司 † 亀井 貴之 † 寺田 純 † 安川 英樹 † 天野 英晴 †

† 慶應義塾大学 理工学部

中規模並列計算機のプロセッサ ·メモリ間結合として MIN が注目されており、近年のプロセッサの急速な高速化にともない、従来よりさらに高い通過率を持つ MIN が提案されている。しかし、従来の MIN は、スイッチングエレメントの構造が複雑であるなど、実装上の問題が多い。我々はこのような実装上の問題を解決する手法として SSS アーキテクチャを提案しており、さらに従来のオメガ網などと比較し高い通過率を持つ PBSF を提唱し、シミュレーションによる評価とともに VLSI への実装を進めている。現在、既にチップのデザインおよびレイアウトが完成しており、本稿ではこの PBSF VLSI チップの詳細について述べる。

相互結合網, VLSI 実装, 並列計算機

PBSF(Piled Banyan Switching Fabrics) Chip based on SSS Network Architecture

M. Sasahara† T. Kamei† J. Terada† H. Yasukawa† H. Amano†

† Keio University

Multistage Interconnection Network (MIN) has been researched for connecting processors and shared memory modules in a middle scale multiprocessor and some new network which has more bandwidth has been introduced for recent high-performace processors.

However, the examples of MIN implementation are quite few because traditional MINs have many troubles with implementation like complicated structure of switching elements.

To cope with this problem, we proposed a novel Multi-stage Interconnection Network (MIN) architecture called Simple Serial Synchronized (SSS)-MIN and high throughput topology called PBSF (Piled Banyan Switching Fabrics) and we've introduced some simulation studies about PBSF's performance. Here, a design and implementation of PBSF VLSI chip is described.

Multistage Interconnection Network, VLSI implementation, Multiprocessor

1　はじめに

Multistage Interconnection Network(MIN) は、数十から数百プロセッサ規模のマルチプロセッサにおけるプロセッサ・メモリ間結合法として注目されてきた [1][2]。

しかし、従来のマルチプロセッサ用に提案された MIN は、スイッチングエレメントの多機能化やピンネック等により、実装に困難な点が多く、並列計算機実装の遅れの一つの原因となってきた。

そこで我々は、パケットをシリアライズし、同期入力にする事によってスイッチングエレメントを単純化し、ハードウェアコストを減少させかつ高速なネットワークを構成できる SSS 型 MIN を提唱し、ネットワークチップの実装を行って来た [3]。

SSS 型 MIN においては、MIN 内での衝突が性能に及ぼす影響が大きいため、高い通過率を持つ結合トポロジが必要になる。そこで我々は複数の出力を持つことにより高い通過率を実現する TBSF(Tandem Banyan Switching Fabrics) の利用、及び PBSF(Piled Banyan Switching Fabrics) の提案を行なってきた。このうち TBSF についてはネットワークチップとともに並列計算機 SNAIL の実装を行い [4]、評価を行って来た。また PBSF については、シミュレーションにより同種の結合網との比較評価を行い、優れた転送能力を持っている事を示した [5]。

現在、我々はこの PBSF-VLSI チップの実装を行っており、既にデザインが完成し、チップ実装の段階にある。本論文では、この PBSF チップの詳細について報告する。

2　SSS 型 MIN

2.1　従来の MIN の問題点

従来マルチプロセッサ用に提案された MIN では、メモリに対するアクセスはパケットの形で 8 ~ 64bit 単位でネットワークに入力される。

ネットワークは Omega 網を代表とするブロッキング網であり、スイッチングエレメント内で衝突が起きた場合にはエレメント内のバッファに格納される。また多くの場合、メモリから読み出されたデータ転送のため、同様の構造を持つ帰還用ネットワークを持ち、プロセッサにデータが渡される。

このような MIN は、複数のプロセッサ間で同時に同一メモリへの参照が集中すると (HotSpot) ネットワーク全体が飽和し、性能が大幅に低下する事が指摘されており、Message Combining[6] に代表される様々な解決手法が提案されている。しかし、これに伴いスイッチングエレメントは多機能・複雑化し、MIN の高密度実装を妨げ、この形式のマルチプロセッサ実装の大きな問題点となってきた。

まず、スイッチングエレメントは、プロセッサあたり 8 ~ 64bit の入力を持つため、LSI チップ内のエレメント数がピン数によって制限される。これによってネットワーク構成のための LSI 数が増え、全体として基板面積の増大を招き、高速なクロックの利用が困難となる。

また、各スイッチングエレメントには、衝突パケットの処理などのために複雑なパケットフロー制御回路とバッファ制御回路が必要になる。さらに、Message Combining 機能を実装するためには高速の多対多のパケット比較を行う並列比較器や Fetch&Add の Combine のための加算器などが必要となる。

2.2　SSS 型 MIN の基本構造

以上のような問題点を解決するため、我々は単純なエレメントから構成され、高い通過率を持つ結合網により、1-4bit 単位で同期入力されたパケットを交換する方式、SSS(Simple Serial Synchronized) 型 MIN を提案している [7]。

SSS 型 MIN の基本構造を図 1に示す。プロセッサからのアクセスは MIN との間のバッファで 1-4bit 程度にシリアル化され、B-ISDN 等で用いられている ATM 用交換システム同様、フレームクロックと呼ばれるクロックに同期してネットワークに入力される。

各スイッチングエレメントは、フレームクロックから生成されるタイミングで入力されたパケットのタグ情報を参照し、それに従ってスイッチを決定する機能しか持たない。さらにスイッチングエレメント内にはパケットバッファは設けられていないため、各エレメントは 1bit 分の記憶しか行わない。従ってスイッチングエレメントの構造は大変単純であり、ネットワークは全体として交換機能を持ったシフトレジスタのように動作する。スイッチに入力された二つのパケットが同一出力に向かった場合、一方のパケットは正しい方向に転送されるが、他方のパケットは正しく転送されず、以後デッドパケットとして扱われる。従ってこのような場合、正しく転送されなかったパケットを次のフレームで再送する必要が

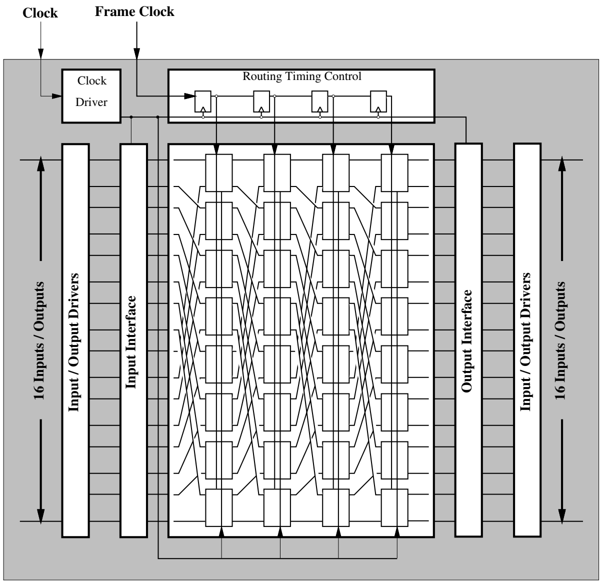

図1: SSS型MINの構造

ある。

ネットワーク内ではフレームクロックの立上り数クロック後1クロック毎に一段ずつスイッチの状態が決定され、段数分のクロック分の遅延の後にネットワーク全スイッチの状態が決定し、ネットワークの入力から出力への論理的なパス(以下トレースと呼ぶ)が設定され、パケットの先頭が出口より出力される。これと同時に、出力側で有効なパケットが到着したかを調べ、ACK(ACKnowledge)もしくはNAK(NotAcKnowledge)をプロセッサ側に返す。プロセッサ側がNAKを受け取った場合、次のフレームでパケットを再送しなくてはならない。

2.3 高速化手法

SSS型MINは基本的にシリアル転送を行うため、ビットパラレル転送の場合と比較し、転送時間の点で不利である。この点を克服するため、シリアル転送・フレーム同期転送の長所を生かした以下のような高速化がなされている。なお、これらの手法はすでに先に実装が行われたSSS型TBSFチップで実装されている。

2.3.1 パイプライン化サーキットスイッチング

SSS型MINでは、アドレス、データ、アクノリッジはの転送路は独立しており、アドレス転送によってトレースが設定されると、アクノリッジおよびデータパケット転送のトレースも決定される。図2に示す通り、これらの転送はオーバーラップして行われる。図中、フレームiでアドレスが転送され、トレースが形成されると、同時に各入力に対してアクノリッジ信号が返される。正しく転送が行われ、ACKが返送された場合、データの転送はこのトレースを利用してフレーム$i+1$でのアドレス転送にオーバラップして行われる(これをパイプライン化サーキットスイッチングという)。

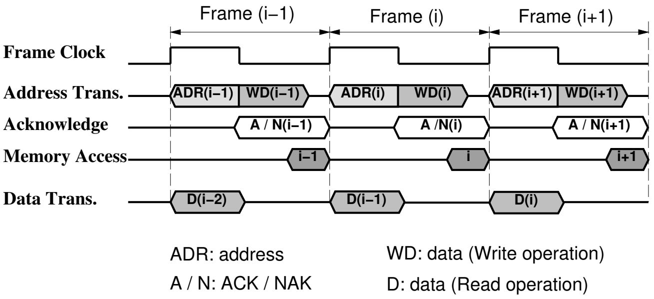

図2: アドレス・データ・アクノリッジの転送タイミング

2.3.2 ステージホッピング

アドレスパケットのルーティングタグにより一度トレースが確定されてしまえば、アドレスパケットの残りとメモリに対する書き込み時のデータは既に設定されたトレースをたどって行けばよい。従って、ルーティングタグ以降のアドレスパケット、データパケット、アクノリッジ信号は各ステージでレジスタに格納されることなく、これらを飛び越してチップ内のマルチプレクサおよび配線による遅延のみで出力に送られる。

2.3.3 メッセージ結合機能

メッセージ結合機能は同一宛先のパケットをエレメント内で結合し、1つのパケットとして扱う事により、Hot Spotへのアクセス集中によるTree Saturationを回避する方法である。この機能の効果はシミュレーション等で確認されているが、従来型のMINでこの機能を実現するためには6倍から32倍のハードウェアを要するため[6]、実現が困難であった。

比べて、SSS型MINではシリアル転送の特徴を生かして簡単に実現できる。SSS型MINのスイッチングエレメントはパケットバッファを持たないため、コンバインができるかどうかは入力されたパケット2つのみを比

較すればよい。さらに、パケットはシリアル化されて入力されるため、比較器は数bit幅のものでよい。

図3に、SSS型MINにおけるコンバインの様子を示す。スイッチングエレメントでは、入力からアクティブなパケットが2つ入力されると、パケットの比較はトレースが確定し、さらにパケットのアドレス部全体がネットワークを通過するまで常に比較が行われている。図はスイッチAにおいて衝突が起こった場合の例であるが、アドレスはトレースが生成された時点では比較が終了しないため、一方は正しく、他方は間違って転送が行われる。しかし、アドレス部の転送の後、両パケットのアドレスが完全に等しい場合、データとアクノリッジのトレースはスイッチ内で図のようにバックワードブロードキャストモードにセットされ、アクノリッジ、データが両方に転送される。

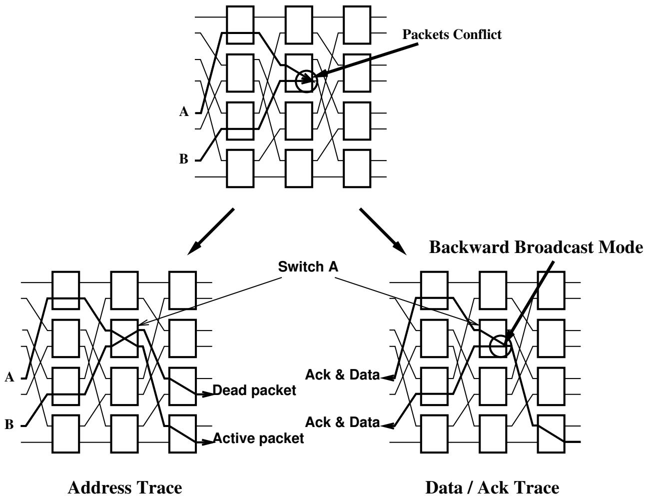

図 3: メッセージコンバイン

さらに、Test&Setのメッセージコンバインも実現されている。あるスイッチングエレメントで衝突したパケットが両方ともTest&Sed操作のパケットで、アドレスが同一でありコンバインが可能な場合、データのスイッチが一方には前段のスイッチからのデータを、他方にはデータに関わらず1(All-1)を次段に送る。これによって、コンバインされた全パケットのうち、1つのみがメモリからのデータを受け取り、他は全てAll-1のデータを受け取る。

2.4 TBSF-VLSIチップとSNAIL

2.4.1 TBSF

このSSS型MINに基づいて我々はTBSF(Tandem Banyan Switching Fabrics)の実装を行ない、これをネットワークとするマルチプロセッサSNAILの実装を行なった。TBSFチップ単体は図1に示す通り通常のオメガ網であるが、これを図4のように直列に接続し、さらに各段の出力をマルチプレクサに入力して構成する。

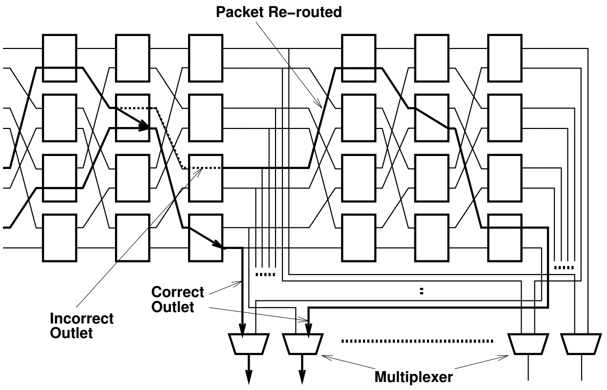

図 4: TBSF(Tandem Banyan Switching Fabrics)

この場合、最初の段のネットワークでパケットが衝突し、あるパケットが正しくルーティングされなかった場合、そのパケットは次段のネットワークに入力され、再度ルーティングされる。このため、衝突によって再送されるパケット数が減るため、ネットワークの負荷が軽くなり、通過率は高くなるが、再度ルーティングされるためレイテンシは大きい。

2.4.2 SNAIL

SNAIL(SSS-Network Architecture ImpLementation)は、16プロセッサ·16メモリモジュールをTBSFで結合したマルチプロセッサである。現在このSNAIL上でTBSFチップの評価を行なっており、開発したTBSFチップが優れた性能を持つことが示されている一方、以下のようなの問題点が上がっている。

- スタベーション

  スイッチ内でパケットが衝突した場合、いずれか一方のパケットのみが正しくルーティングされるが、TBSFチップでは、どちらのパケットが競合

に勝つかが静的に決定されている。このため、各入力に対して静的にプライオリティが決まってしまい、スタベーションを起こす可能性がある。

- 転送能力

  SNAIL のプロセッサはモトローラのMC68040(CISC)であり、さらに転送能力を稼ぐため TBSF チップ4 つを並列に使っているため、現在の構成では、ネットワークは十分な転送能力を持っている。

  しかし、プロセッサを RISC とした場合を想定すると、段数を増加させて通過率を改善しても、2 段目 3 段目通過にともなう遅延が大きく、プロセッサのパフォーマンスを低下させてしまうことが予測される。

スタベーションは SSS 型 MIN 自身の問題ではなく、実装上の問題であり、回避方法も多々考えられるが、転送能力の問題は重大である。そこで、我々はレイテンシを低く抑えつつ、かつ高い通過率を持つネットワークとして PBSF チップの実装を行なった。

3 PBSF(Piled Banyan Switching Fabrics)

PBSF は図 5に示すように banyan(omega) 網を三次元的に接続した構造を持つ [7]。図のように最上層と最下層を除く層のスイッチングエレメントは水平方向の入出力を 2 つずつ、 垂直方向を 2 つずつ、 計 4 入力 4 出力を持つ。パケットは水平方向に進み衝突が起こって進めなくなると、SSS 型 MIN では衝突によりルーティングに失敗したパケットはデッドパケットとなるが、このPBSF では、 正しくルートされなかったパケットは下の層のネットワークに送られる。

パケットはまず最上層のネットワークに入力される。最上層において、 水平方向に進む 2 つのパケットがあるエレメントで出力リンクの競合が起こると、片方のパケットは希望の方向に送られ、 競合に敗れたパケットは一つ下の層のエレメントに送られる。

2 層目以下では、 水平方向同士の衝突に加え、更に水平方向に進んで来るパケットと上層から送られて来たパケットとが競合する。 この場合、 最大で 3 つのパケットが一つの出力リンクを競合する (水平方向からのパケット 2 つと垂直方向からのパケット 1 つ) が、この時も基本的には最上層のスイッチの動作と同様で、それらのパケットのうち一つは正しい出力へと送られ、 もう一つは更に下層へ、残りの一つはスイッチングエレメント内で消滅する。最下層で 3 つのパケットが競合した場合、 下層へ送る事ができないため、 3 つのパケットのうち一つだけが正しく転送され、残りのパケットはエレメント内で消滅する。

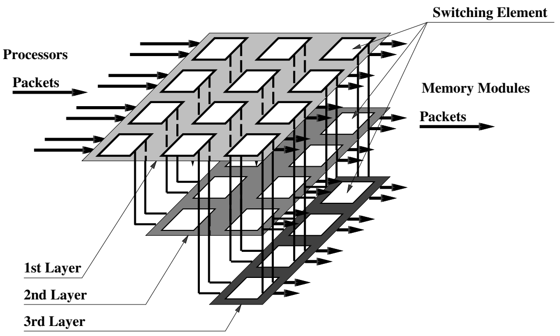

図 5: PBSF (Piled Banyan Switching Fabrics)

いずれの場合も、 パケットが消滅した場合にはプロセッサ側に NAK が返されるようになっており、消滅したパケットはネットワークインタフェースによって次のフレームで再送される。

PBSF において、下層の網は、従来型の MIN がスイッチングエレメント内に持つパケットバッファと同様の効果がある。このことから、層を増やす事によって通過率、通過時間両方の改善が期待できる。

4 実装

4.1 予備評価

図 5に示したように、PBSF はフルサイズでオメガ網の段数分 (logN) のレイヤーを持ち、規模の大きいネットワークを実装する際にはスイッチングエレメントのハードウェアコストが問題になる可能性がある。しかしPBSF ではその構成上、 2 層目以下の層やスイッチをある程度省いても問題はなく、実装の際の条件にあわせ実装を行うことができる。

また、SSS 型 MIN 自身、アドレスパケットおよびデータのビット幅も任意であり、これも同様に実装の条件に応じて実装を行う事になる。

そこで、ここで実装に際してネットワークの規模や付加機能等の詳細を定めるために行った予備評価について

述べる。

評価はWS上でPBSFのC言語によるシミュレータを作成し、今回実装を行うPBSFと同一の規模、すなわち16入力16出力の場合を想定して行った。

4.1.1　レイヤ数

PBSFのハードウェア量を最も大きく左右するのはレイヤを何層設けるかである。図6に、ランダムパケットをネットワークに入力した際の通過率を示す。図から分かる通り、レイヤー数が1、すなわち通常のオメガ網からレイヤー数を増加させる従い、通過率が改善される。しかし、レイヤー数が2から3に増加した場合、通過率の伸びはそれほど大きくはなく、3から4に増やした場合、通過率は多少は改善されるに留まる。このことから、16入出力程度の規模では、2レイヤー程度で十分である事がわかる。

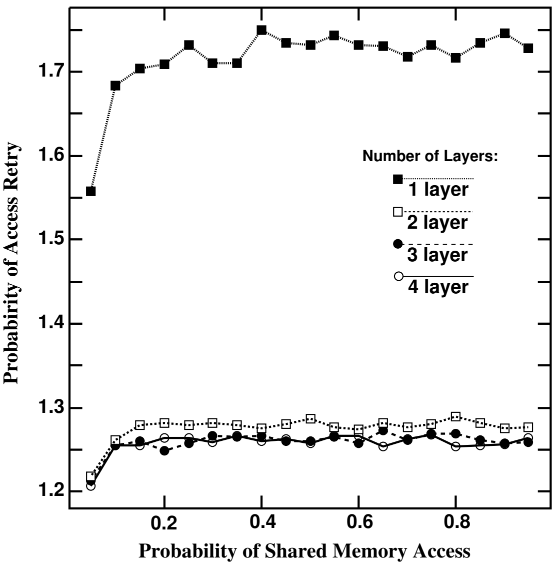

図6: レイヤ数に対するネットワーク通過率

4.1.2　優先度制御

先に述べた通り、SNAILに用いられているTBSFチップではスタベーションが起こる可能性があった。今回実装するPBSFでは、これを回避するため、パケットにプライオリティを付加し、優先度制御を行なうこととした。

プライオリティはスイッチの状態を決める際にパラレルデータとして各スイッチに伝達されていなくてはならないため、チップ内の配線量の増大を招く。そこで、実装時にレベル数を何レベル設けるかを決定するため、プライオリティの効果について評価を行った。

図7はPBSFにプライオリティ制御機能を付加し、負荷1で5000ネットワークフレーム分入力し、各入力のパケットが最大何回の再送が行われたかを記録したものである。ただし、プライオリティは、最初にパケットを転送する際にはパケットの優先順位を最低にし、再送を繰り返すたびにプライオリティを上げるという方法で決定されている。図は横軸にプロセッサID、縦軸が最大再送回数を示している。従って、グラフのプロットが上下にばらつき、高い位置にあるほどスタベーションの影響が大きいことを示す。

図中、プライオリティがない場合は各入力間で最大数のばらつきが大きく、スタベーションの影響を顕著に現しており、また再送の最大回数も大きい。しかし、プライオリティのレベル数を増加させるに従い、各入力間の再送回数のばらつきは小さくなり、また再送の最大回数も減少する。

この現象はとくにプライオリティが2レベル(1bit)から4レベル(2bit)となる時が顕著であり、4レベルの優先順位を設ける事で、2レベル時は最大で10回以上行なわれた再送の回数を5~6回程度におさえることができることがわかる。また、レベル数を4レベルから8レベル(3bit)、16レベル(4bit)に増加させてもグラフに大きな変化はなく、再送回数も4回程度にとどまる。このことより、プライオリティの割り当て方にも依存するが、プライオリティによるパケットフローコントロールは効果が大きく、今回の規模のネットワークであれば4レベル程度のプライオリティを設けることによってスタベーションは回避できると予想される。

4.2　PBSF VLSIチップ

実装されたVLSIチップは、16入力16出力、2レイヤで構成されるPBSFが1チップに実装されている。各入力はアドレス2bit、データ1bit幅を持ち、50MHz動作時で1入力あたり100Mbit/sの転送能力を持つ。また、最上層のレイヤにはメッセージコンバイン機能が実装され、さらに4レベルのパケットのプライオリティ制御をサポートする。

今回のチップの作成は電子情報通信学会VLSI設計教育高度化専門委員会のVLSI試作試行の一環として

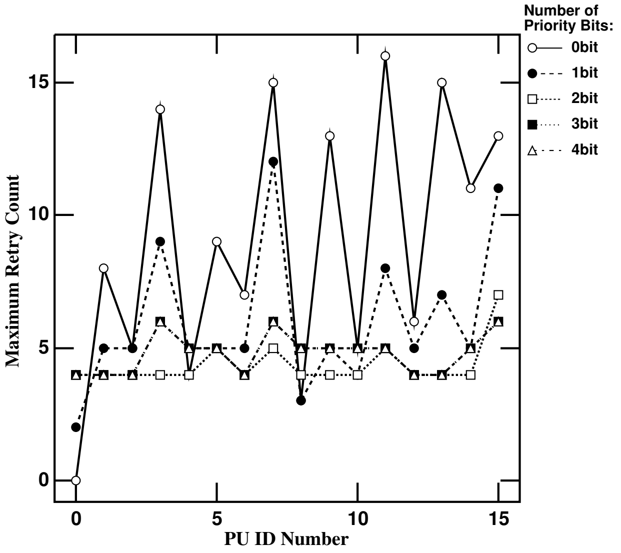

図 7: プライオリティのレベル数に対する最大再送の回数

行った。

このため、実装に当たって使用するチップのパッケージおよびテクノロジ、ゲート数は設計に先立って決定されており、PBSF 設計に際しての詳細な仕様はこれらの制約と先にのべたシミュレーションの結果から決定された。

- ビット幅

  信号ピン数が 116 ピンであるため、16 入出力とした場合、各チャネルに割り当てられるピン数は 7 本である。これより割当をアドレス 2bit ・データ 1bit 幅 (入力 ・出力で計 6 本)、およびアクノリッジとした。また、このため各レイヤーの出力を個別に出力せず、上下のレイヤの出力をマルチプレクスした後出力するようにした (図 8)。

  しかし、チップの内部では、ルーティングを行なう際にスイッチングエレメントに与えるデータとして、ルーティングタグ、パケットのアクティビティ、プライオリティ(2bit) の計 4 本の信号が必要であり、これらはアドレス転送用のワイヤを使ってパラレルにスイッチングエレメントに渡さるため、今回のチップでアドレスを 4bit パラレルとしても、ハードウェア量はほとんど変化しない。今回の実装に用いられたパッケージは決して大きいものではなく、またアドレスを 4bit 幅としてもピン数の増加は 30 本程度で、200 ピン程度の PGA で実装可能であることを考えると、大変残念である。

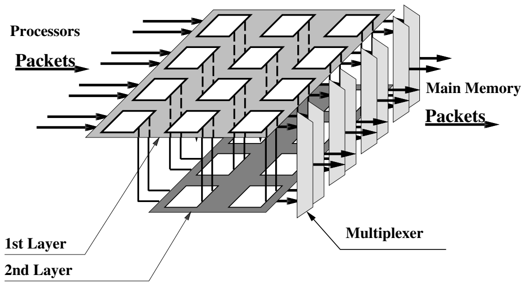

図 8: 実装された PBSF(ただし 図は 8 入出力の例)

- メッセージコンバイン機能

  SSS 型 MIN では、基本的にメッセージコンバインを実装する際のハードウェア量は大変小さい。しかし、今回の実装では全スイッチにメッセージコンバイン機能を付加した場合、わずかに収容可能ゲート数を越える恐れがあり、デザイン提出が遅れる可能性があったため、メッセージコンバインは最上層と最後のマルチプレクサでのみ行なうこととした。

PBSF チップの仕様を表 1に示す。

表 1: SSS-PBSF チップの仕様

Clock	100MHz
Bit width	2bit(address) 1bit(data)
Size	16-in/out
Max bandwidth	$300 Mbits/sec \times 16$
セル使用数	17356
信号ピン数	116
テクノロジ	$0.5\mu m$ CMOS sea-of-gates

参考までに、先に実装された SSS-TBSF チップ (これは通常のオメガ網と同一である) の仕様を表 2に示す。

いずれも 2×2 のスイッチを使っており、また、先に述べた通り今回実装した PBSF では、アドレスは内部で 4bit 幅分の配線量を持っているため、両者がスイッチ構成に要するゲート数 ・配線量はほぼ同一と考えられ

る。従って、PBSF は TBSF と比較し、56/32 = 1.75 倍のハードウェアを必要とすることが予測される。

しかし実際には、レイヤー数を増加させ、かつ 4 レベルのパケットプライオリティを付加してもゲート数の増加は 17356/9838 = 1.76 倍に留まっており、SSS 型 MIN が拡張性に優れていることがわかる。

表 2: SSS-TBSF チップの仕様

Clock	50MHz
Bit width	3bit(address) 2bit(data)
Size	16-in/out
Max bandwidth	250$Mbits/sec \times 16$
セル使用数	9838(セル使用率 49%)
信号ピン数	180
テクノロジ	1.0μm CMOS sea-of-gates

なお、このチップの設計は Verilog-HDL で行なった。

5 おわりに

本稿では、SSS 型 MIN の構造、PBSF の構成および実装された PBSF VLSI チップの詳細について述べた。SSS アーキテクチャの応用により、高い通過率を持ち、ハードウェアの負担の少ない PBSF を高い実装密度で効率良く実装できることが示された。

今回の実装では、パッケージの大きさやゲート数が十分ではなく、残念であるが、今後このネットワークチップを用いたマルチプロセッサを実装し、実機上で PBSF の実アプリケーションでの耐混雑性などについて評価を行なっていきたい。

参考文献

[1] J.Konicek, et al,: *The Organization of the Cedar System*, Proc. of 1991 Int. Conf. Parallel Processing, (1991).

[2] G.F.Phister, et al.: *The IBM Research Parallel Processor Prototype (RP3): Introduction and architecture*, Proc. of 1985 Int. Conf. Parallel Processing, (1985).

[3] 周洛, 天野英晴: SSS-MIN 用プロトタイプチップの実装, 情報処理学会第 44 回全国大会論文集,6-103,104 (1992).

[4] M.Sasahara, et al: *SNAIL:a multiprocessor based on the Simple Serial Synchronized multistage interconnection network architecture*, Proc. of International Conference on Parallel Processing, pp.I117-120, Aug. 1994

[5] T.Hanawa, H.Amano, Y.Fujikawa: *Multistage Interconnection Networks with multiple outlets* Proc. of International Conference on Parallel Processing, pp.I1-8, Aug. 1994

[6] G.F.Phister and V.A.Norton : *Hot Spot Contention and Combining in Multistage Interconnection Networks*, IEEE Trans. on Comput. vol. c-34, No.10, (1985).

[7] 天野, 周, 藤川: SSS (Simple Serial Synchronized) 型マルチステージネットワーク, 情報処理学会論文誌 Vol.34, No.5, pp. 1134-1143, (1993).