1 はじめに

多段結合網 (Multistage Interconnection Network: MIN)[3] は中規模 (数十から数百プロセッサ) の並列計算機におけるプロセッサとメモリ間の接続に多く用いられる結合方式である. 代表的な MIN では, メモリに対するアクセスは, パケットの形でスイッチングエレメント間をビットパラレル (8 ~ 64bit) に転送する. 結合網のトポロジは Omega 網を代表とするブロッキング網で, スイッチングエレメント内で衝突が起きた場合, 片方のパケットはエレメント内のバッファに格納される. メモリから読み出したデータをプロセッサ側に送るため, 同様の構造を持つ帰還用の MIN が必要になる.

このような従来型の MIN は, 高速, 高密度実装に困難な点が多く, これを用いたマルチプロセッサシステムの実装の遅れ, 規模縮小 [5][9] の原因の 1 つになってきた. まず, プロセッサ当たり 8 ~ 64bit の入出力線を必要とするため,LSI チップ中に実装できるエレメント数がピン数の制約によって少なくなる. このことは, 全体として用いるチップ数, 基板面積の増大を招き, 高速なクロックの利用を難しくする. さらに, エレメント構造と制御が複雑である. パケット同士が衝突した場合, 内部のバッファに格納するため, 前後のスイッチングエレメントとの間にハンドシェークとバッファの溢れ制御が必要となる.

これらの問題点の解決のため, 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, ビットシリアルに同期して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 [12][17]. このことにより, 高速なクロックの利用を可能とすると共に,LSI のピン数制約の問題を解決し, 高密度実装を可能とする.

しかし一方で SSS 型 MIN はエレメント内にバッファを持たないためパケット同士の衝突が起きた場合, 入力バッファから再送する必要があり, 性能低下の大きな原因となる. このため,SSS 型 MIN で用いる結合網は従来型 MIN で用いられるブロッキング網よりも高い通過率が要求される. 一般に通過率の高い網はスイッチングエレメントを数多く必要とするが, SSS 型 MIN ではスイッチングエレメントの構造が簡単であるため, ある程度の数は許容できる. クロスバや Clos 網 [1] に代表されるノンブロッキング網は、ブロッキング網よりも高い通過率を持つが、それでもサイズがある程度以上になると出力が均等に分布した場合は、67%前後に制限されてしまう. これは、同一出力に向かうパケットが出力線で競合するためである.

そこで、我々は、複数の出力を持つことにより、同一出力に向かう複数パケットの通過を許す MIN について検討した. まず、MIN の縦列接続から成る TBSF(Tandem Banyan Switching Fabrics) を検討し [12], さらにこれを改良し三次元的な接続にした PBSF(Piled Banyan Switching Fabrics)[16] を提案した. また、類似の MIN として複数の独立の MIN を並列に使う Multi-Banyan[8]、倍のサイズを用いる Double-Banyan[14] が知られている. 本論文では、これらの MIN の通過率を理論的に解析し、比較検討する. また、行列をアクセスする場合の能力に関しても理論的に検討し、比較する.

2 SSS 型 MIN

多重出力可能な MIN は、B-ISDN 用の高性能パケット交換網への応用も考えられるが、ここでは、SSS 型 MIN としてプロセッサ-メモリ間接続に用いる場合を主に考える. そこで、まず、SSS 型 MIN について MIN の制御に関連した部分を簡単に紹介する. SSS 型 MIN の構造と動作の詳細は [12][17] に示されているので参照されたい.

2.1 SSS 型 MIN の基本動作

SSS 型 MIN において, メモリアクセスに相当するパケットは, プロセッサとの間に接続された入力バッファから, パケット同期信号 (フレーム信号) に同期してビットシリアルに MIN へ入力される. 結合網は 2×2 のスイッチングエレメント多数から成る通過率の高い網を用いる. 各エレメントの構造は非常に単純であり, 基本的にはパケットの 1bit 分の記憶のみを行なう.

各段のスイッチングエレメントは, プロセッサからメモリモジュールへ向かうアドレス線, メモリモジュールからプロセッサに向かう応答信号線, 双方向のデータ線の 3 本の信号線によって, 互いに結合される. プロセッサは, ルーティングタグとモジュール内アドレスから構成されるアドレスパケットを, フレーム信号に同期してアドレス線に対して送る. ルーティングタグは, 行き先のメモリモジュールの番号と,2bit の制御ビット (data valid, conflict bit) から成る.

従来型 MIN 同様, スイッチの状態はルーティングタグ内のモジュール番号に従って, 各エレメント内で独立に決定される. 2 つの入力パケットがエレメントの同一出口に向かった場合, 衝突が発生する. この場合, 従来型 MIN とは異なり SSS 型では各エレメントが基本的にはパケットの 1bit しか記憶できないため, 衝突を起こしたパケットのうち 1 つは, 希望しない方向に送られる. この時, 希望の方向に進めなかったパケットの conflict bit がセットされる. 以降, この bit がセットされたパケットはダメージパケットと呼ばれ, 他のパケットの進行を妨害しない.

アドレスパケットの先頭が MIN の出口 (メモリ側) に達した時, すべてのエレメントの状態が決定され, 入出力間に論理的なパスが設定される. このパスをトレースと呼ぶ. パケットは衝突によって, 正しくないモジュールに送られる可能性がある. このため, アドレスパケットの先頭が MIN の出口に達した時にルーティングタグ内の conflict bit をチェックし, 正しく到着したかどうかを示す応答信号 (ACK:正しく到着,NAK:衝突により希望しないモジュールに到着) を, 応答信号線を用いて即座にプロセッサ側に返す. この応答信号は, 設定されたトレースに従って, エレメント内で記憶されずに, マルチプレクサと信号線の遅延だけで伝搬する. NAK を受けとった場合, 同じパケットが, 次のフレームで再び入力バッファから SSS 型 MIN に対して送られる.

アドレスパケットにより設定されたトレースは, データ線に関しては 1 フレーム遅れて有効になる. このことにより,i 番目のフレームで設定されたアドレスに対するデータ転送は,i+1 番目のフレームで, 次のアドレスパケット転送とオーバラップして行なわれる. トレースの共有により, データパケットの制御はアドレス用の制御信号を 1 フレーム遅延させてそのまま用いることができる. この方法により,SSS 型 MIN では帰還用の MIN が必要なくなる.

さらに、アドレスパケット転送中にその内容をエレメント内で比較し、一致した場合、データのトレースを Tree 状に設定していくことにより、簡単にメッセージ Combine が可能である.

3 多重出力可能な結合網

以下、本論文で検討する多重出力可能な結合網に関して順に解説する.

3.1 TBSF(Tandem Banyan Switching Fabrics)

TBSF は, 本来 B-ISDN(Broadband Integrated System DigitalNetwork) で用いられる ATM(Asynchronous Transfer Mode) パケット交換用の網として、国内では我々と沖電気の共同研究により 1988 年 [7] に、海外では Tobagi らにより 1990 年 [10] に、提案された網である. ATM 網用 [11]、SSS 型 MIN 用 [15] にそれぞれこの網専用の LSI の実装が行なわている.

TBSF は, 図 1に示すように banyan 網 (omega 網) を直列に接続し, 各網の出口にバイパス路を設けた構造を持つ. 宛先に到着したパケットはバイパス路によりそのままメモリモジュールに送られ, 衝突により到着しなかったパケットのみが次の段の banyan 網に入力される.

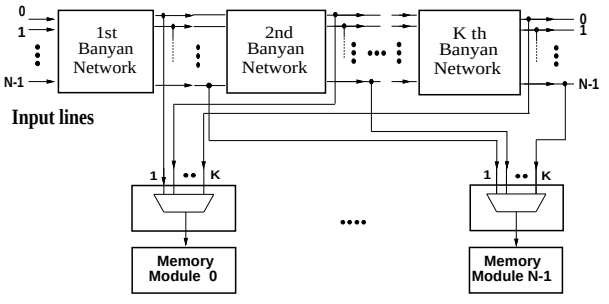

図 1: Tandem Banyan Switching Fabrics の構成

ある箇所でセットされた conflict bit は, 次の網の入力でリセットされ, 再びルーティングが行なわれる.

TBSF は B-ISDN 用の交換機に用いる場合, 各 banyan 網から時間的にずれて到着するパケットを揃える部分が問題になる. しかし,SSS 型 MIN として用いる場合, 到着順にメモリのアクセスを行えばよいので, メモリ利用率を上げることができ, 利点に転じる. 1 フレーム内に何回メモリのアクセスが許されるかは, メモリモジュールのアクセス時間とフレーム時間により定まるが, 通常 2,3 回が限度である. それ以上到着したパケットに対しては衝突時同様に NAK 信号が返され, 次のフレームで再送が行なわれる.

モジュラリティに優れ, 耐故障性にも優れた点を持つ [11].

3.2 PBSF(Piled Banyan Switching Fabrics)

TBSF では、接続された各網において、パケットが衝突するまでは正しくルーチングされるにもかかわらず、それまでのルーチングの結果は、次の網に対して全く貢献しない. この点を改良するため、我々は、図 2に示すように banyan(omega) 網を三次元的に接続した構造に変更し、PBSF(Piled Banyan Switching Fabrics)[16] と名付けた. 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.

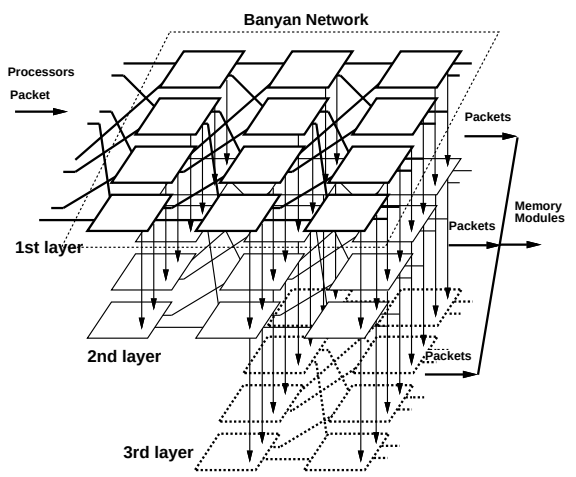

図 2: Piled Banyan Switching Fabrics の構成

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の出力リンクで競合すると, 片方のパケットは希望の方向に送られ, 敗れたパケットはひとつ下の層のエレメントに,1 クロック (厳密には半クロック) の遅延を伴って送られる. 上層から送られたパケットは下の階層を水平方向に進んでくるパケットと競合し, 上層からのパケットが優先的に送られ、敗れたパケットはさらに下の階層に,1 クロックの遅延の後に送られる. 水平方向からのパケット 2 つ, 垂直方向 (上層) からのパケットがすべて同一出力線を目指した場合, 上層からのパケットが目的の出力線に進み, 水平方向からのパケットのうちのひとつが選ばれ, 下層に送られる. 残った 1 つのパケットは conflict bit がセットされ, 希望しない方の出力に送られる. 最下層のネットワークで競合に敗れた場合は衝突を示す conflict bit がセットされ, 以後他のパケットを妨害しない.

TBSF ではフレーム時間分ずらして次の網に送ったパケットを PBSF では 1 クロック分だけずらして, 下の層の網に送る. 下層の網は, 従来型の MIN が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. このことにより, 通過率, 通過時間両方の改善が期待できる.

網の通過時間を経過した後,PBSF の各層からパケットが出力される. 網の出力からは到着できなかったパケットが出力される可能性があり, conflict bit のチェックが行なわれ, トレースを用いた応答機構により入力バッファに転送の失敗が通知され, パケットの再送が行なわれる. TBSF 同様メモリシステムには層数分のパケットが到着する可能性がある. TBSF とは異なり, これらのパケットはほぼ同時に到着するので, メモリシステム内にバッファを設け, これらを蓄えて順にアクセスする. アクセス回数が多くフレーム時間内にすべてのデータを返せない場合, 次のフレームで到着するパケットが失敗と見なされ, 再送される.

3.3 MBSF(Multi Banyan Switching Fabrics)

独立の Banyan 網を複数用意し、トラフィックを分散する方法は、パケット交換網、サーキットスイッチ用の交換網などに、古くから検討が行なわれてきた [4][6][8][13]. ここでは、この網を他の網にならって、MBSF(Multi Banyan Switching Fabrics) と呼ぶ. この場合、並列に接続された網へのパケットの配分方法が問題になるが、ここでは、ランダムに割り付けると仮定した. TBSF、PBSF においては複数の Banyan 網がお互いにパケットのトラフィックに何らかの影響を与えるが、MBSF ではこれらは全く独立しており、単に各網に対する負荷が網数分に分散される.

3.4 DBSF(Double sized Banyan Switching Fabrics)

Banyan 網のサイズを倍にして、その半分の入力に対してパケットを与えると、負荷に関して MBSF とほぼ同等の効果が得られることが期待される. この方法は複数パケットの通過を許すわけではないが、衝突の状況は類似すると考えられ、すでに Lawrie によって配列に対するアクセス能力が解析されている [2]. そこで、本論文では、複数パケットの通過を許す効果の比較対象として評価することにした. 他の網にならって DBSF(Double sized Banyan Switching Fabrics) と呼ぶ.

3.5 定性的な評価

TBSF,PBSF,MBSF,DBSF をまず、通過率、レイテンシィ、コストの点で、定性的に検討する.

通過率 複数のパケットの通過を許す点で TBSF,PBSF,MBSF は DBSF より高い通過率が期待できる. さらに、MBSF は最初にトラフィックを分離してしまうため、網の性能を十分引き出すことができず、使用網数が多くなれば、TBSF,PBSF に比べ不利である. また、PBSF は、衝突前のルーティングが有効になるため,TBSF より若干高い通過率が期待できる.

レイテンシィ 縦列接続を行なう TBSF が最も不利である. ここでは、入力数を N , 縦列または三次元方向に接続する網数を K とする. まず、MBSF は通常の Banyan 網と変わりがないので、通過時間は $log_2 N$となり、DBSF はそれより 1 大きく $log_2 N + 1$ である. しかし、TBSF の通過時間は $K \times log_2 N$,PBSF は $K - 1 + log_2 N$ である. すなわち,K の数がさほど大きくない場合、レイテンシィは MBSF, DBSF, PBSF はほぼ同等で、TBSF に比べ, ほぼ $1/K$ になる.

コスト まず、エレメントのハードウェアについては、TBSF, DBSF, MBSF は通常の Banyan 網と同じである. これに対し、PBSF のスイッチングエレメントは垂直方向の入出力が加わり,4 入力 4 出力である. このエレメントは, 垂直方向の線が, エレメントの出力に対して設けられ, 優先的に通路を確保する構造を取る. このため, ハードウェア量は通常の 4×4 のクロスバより小さい. しかし, それでも TBSF のスイッチングエレメントの約 1.5 ~ 1.8 倍程度のハードウェアを必要とする.

一方、エレメント数は DBSF が $N(log_2 N + 1)$ になる以外は、$(KN)/2log_2 N$ となり、等しい. すなわち $K = 2$ の時のみ DBSF がやや大きいが、それ以上では他の結合網が大きくなる.

4 解析モデル

本章では、前章で述べた各結合網について、解析モデルとして表し、その評価を行う. 均一なトラフィックを仮定した場合の SSS 型 MIN の通過率は、Wu らによるサーキットスイッチングに基づく MIN の解析手法 [13] を応用することによって解析することができる.

本論文では次のことを仮定する.

1. 各入力へのパケットの入力確率は均一
2. パケットの行き先のアドレスは等確率に分散

4.1 TBSF

まず、banyan 網の一段目について考える.

トラフィックが均等であり、各エレメントの入力にパケットが入力される確率を r とすると、両方の入力にパケットが入力され、かつ行き先が同一である時のみ衝突が生ずる. 従って、衝突の起こる確率は $0.25r^2$ である. ここでは、衝突によって競合に敗れたパケットはダメージパケットとなり、他のパケットを妨害しないため、消失したものと見做すことができる. 従って i ステージ目のエレメントの入力にパケットが入力される確率を r_i とすると、

$$r_{i+1} = r_i - 0.25{r_i}^2 \equiv X(r_i) \tag{1}$$

となる. ここで、$X(r)$ とは r の確率でエレメントにパケットが入力されたときパケットが出力される確率を表す. banyan 網一段において最終的にパケットが出力される確率は全ステージ数 n とすると

$$r_n = X^n(r_0) \tag{2}$$

となる. ここで r_0 は、banyan 網の各入力にパケットが入力される確率 (負荷率) である.

TBSF では、banyan 網を衝突せずに通過したパケットはメモリモジュールに送られるため、次の banyan 網ではそのパケットは除かれる. そこで、k 番目の banyan 網の各入力にパケットが入力される確率を R_{k-1} とすると、次の式が成立する.

$$R_k = R_{k-1} - r_n(R_{k-1}) \tag{3}$$

ここで $r_n(R_k)$ は R_k の初期値により (2) 式を計算したものである. さらに k 番目の banyan 網の通過率は

$$R_k/R_0 = \{R_{k-1} - r_n(R_{k-1})\}/R_0 \tag{4}$$

で表される.R_0 は TBSF の各入力にパケットが入力される確率である. 接続段数 l 段の TBSF 全体の通過率は 各 banyan 網における通過率の総和 $\left(\sum_{k=1}^{l} R_k\right)/R_0$ となる.

4.2 PBSF

PBSF では、他の結合網とは異なり、競合が起こるとパケットは垂直方向に送られるため、垂直方向のパスについても考える必要がある.

まず、banyan 網の一段目は、垂直方向の入力はないので、TBSF の時と同様に考えることができる. 従ってステージ i での水平方向の出力は (1) 式と等しくなる. また、垂直方向に出力されるのは スイッチングエレメント内で衝突する場合であるから、

$$d_{i,2} = 0.25{r_i}^2 \tag{5}$$

となる. ここで $d_{i,2}$ とは 2 段目の banyan 網のステージ i での垂直方向の入力を表す.

つぎに 二段目以降の banyan 網について考える. 各エレメントへの入力パターンを考え、各状態の生起確率と その場合の水平または垂直方向への出力確率を列挙すると 表 1のようになる.

表 1: スイッチングエレメントのパケットの出力確率 (PBSF)

水平入力		垂直入力		生起確率	水平出力		垂直出力	
$1-d$	$1-d$	$1-r$	r	$(1-d)^2 r(1-r)$	0.5	0.5	0	0
$1-d$	$1-d$	r	$1-r$	$(1-d)^2 r(1-r)$	0.5	0.5	0	0
$1-d$	$1-d$	r	r	$(1-d)^2 r^2$	0.75	0.75	0.25	0.25
$1-d$	d	$1-r$	$1-r$	$d(1-d)(1-r)^2$	0	1.0	0	0
d	$1-d$	$1-r$	$1-r$	$d(1-d)(1-r)^2$	1.0	0	0	0
d	$1-d$	$1-r$	r	$d(1-d)r(1-r)$	1.0	0.5	0.5	0
$1-d$	d	$1-r$	r	$d(1-d)r(1-r)$	0.5	1.0	0	0.5
d	$1-d$	r	$1-r$	$d(1-d)r(1-r)$	1.0	0.5	0.5	0
$1-d$	d	r	$1-r$	$d(1-d)r(1-r)$	0.5	1.0	0	0.5
$1-d$	d	r	r	$d(1-d)r^2$	0.75	1.0	0.25	0.75
d	$1-d$	r	r	$d(1-d)r^2$	1.0	0.75	0.75	0.25

従って、k 番目の banyan 網のステージ i のエレメントにおいて、パケットが出力される確率は

$$\begin{aligned} r_{i+1} &= (1-d_{i,k})^2 r(1-r_i) + 0.75(1-d_{i,k})^2 {r_i}^2 \\ &\quad + d_{i,k}(1-d_{i,k})(1-r_i)^2 \\ &\quad + 3d_{i,k}(1-d_{i,k})r_i(1-r_i) \\ &\quad + 1.75 d_{i,k}(1-d_{i,k}){r_i}^2 \equiv X(r_i, d_{i,k}) \qquad (6) \\ d_{i,k+1} &= 0.25(1-d_{i,k})^2 {r_i}^2 + d_{i,k}(1-d_{i,k})r_i(1-r_i) \\ &\quad + d_{i,k}(1-d_{i,k}){r_i}^2 \equiv Y(r_i, d_{i,k}) \qquad (7) \end{aligned}$$

となる. 従って、k 番目の banyan 網の出力は、banyan 網の入力を r_0 、全ステージ数 n とすると

$$\begin{aligned} r_n &= X(r_{n-1}, d_{n,k-1}) \\ &= X(X(r_{n-2}, d_{n-1,k-1}), Y(r_{n-1}, d_{n,k-2})) \\ &\vdots \end{aligned}$$

となる. そこで、k 番目の banyan 網の出力を R_k とすると、次の式が成立する.

$$R_k = R_{k-1} - r_n(R_0, R_1, \ldots, R_{k-1}) \tag{8}$$

ここで、R_0 は PBSF の各入力にパケットが入力される確率である. 従って接続段数 l 段の PBSF 全体の通過率は TBSF 同様、各 banyan 網における通過率の総和となる.

4.3 MBSF

この結合網は、各 banyan 網は独立であり、banyan 網同士で依存関係がないため、モデル化は非常に容易である.

まず、MBSF では、各 banyan 網にパケットを分散させるために、最初のステージにデマルチプレクサを用意し、ランダムに各 banyan

網にパケットを振り分けている. 従って、パケットが入力される確率を r_0 、banyan 網の数を l とすると、各 banyan 網にパケットが入力される確率 r_1 は

$$r_1 = r_0/l \tag{9}$$

となる. 各 banyan 網の下では、均一なトラフィックの下で、各スイッチングエレメントへのパケットの入力確率を r としたとき、そのエレメントからパケットが出力される確率は (1) 式の通りである. k 番目の banyan 網の出力は、banyan 網の入力 r_1、banyan 網のステージ数を n とすると、

$$r_{n+1} = X^n(r_1) \tag{10}$$

となる. そこで k 番目の banyan 網の出力を R_k とすると、次の式が成立する.

$$R_k = r_{n+1}(R_0/l) \tag{11}$$

ここで、R_0 とは、MBSF の各入力へのパケットの入力される確率を示している. 接続段数 l 段の MBSF 全体の通過率は TBSF,PBSF 同様、各 banyan 網における通過率の総和となる.

4.4 DBSF

まず、最初のステージでは、DBSF の入力が N なのに対し、エレメントの入力が $2N$ であるので、各エレメントの入力へパケットが入力される確率は

$$r_1 = r_0/2 \tag{12}$$

である. ここで r_i はステージ i の入力 (ステージ $i-1$ の出力) を表す. ステージ 2 から ステージ $n-1$ (n は DBSF の全ステージ数) のスイッチングエレメントの場合は、(1) 式と同様である. ステージ n (最終ステージ) では、二つの入力からの OR になるが、両方の入力にパケットが入力された場合には片方のパケットは失われる. 従って、最終ステージでパケットが出力される確率は

$$r_n = 2r_{n-1} - {r_{n-1}}^2 \tag{13}$$

となる.

5 シミュレーションと評価

前章では、多重出力可能な MIN の解析モデルについて検討した. 本章では、これらの MIN をシミュレーションを用いて通過率を測定し、解析モデルの正当性と、各 MIN の通過率について評価を行う.

5.1 シミュレーション

シミュレーションは各結合網に対して 10000 回行った. 各結合網は 256 入力を仮定した.

図 3 に TBSF, 図 4 に PBSF, 図 5 に MBSF の、banyan 網の段数を 1-5 としたときの負荷率と通過率の関係、図 6 に DBSF の負荷率と通過率の関係を示す.

このグラフから、解析結果とシミュレーション結果との差は高々 5%程度であり、解析のためのモデル化が適当であることを示している.

次に ある一定の負荷率で banyan 網の段数を 1 ~ 5 段に変化させた時の通過率の変化を 図 7、8に示す. 図 7は負荷率 30%、図 8は負荷率 80% の場合を表す.

これらのグラフから TBSF と PBSF は、MBSF,DBSF に比べて通過率が非常に高く、banyan 網数 3 段程度で 90%近い通過率が得られることがわかる.

図 7 のように網の負荷率が低い時には TBSF,PBSF と MBSF や DBSF の通過率の差は、さほど大きくない. しかし、図 8 のように負荷率が高くなると、TBSF,PBSF の通過率はあまり低下しないが、MBSF,DBSF の通過率は大きく低下し、通過率の差は大きくなっていく.

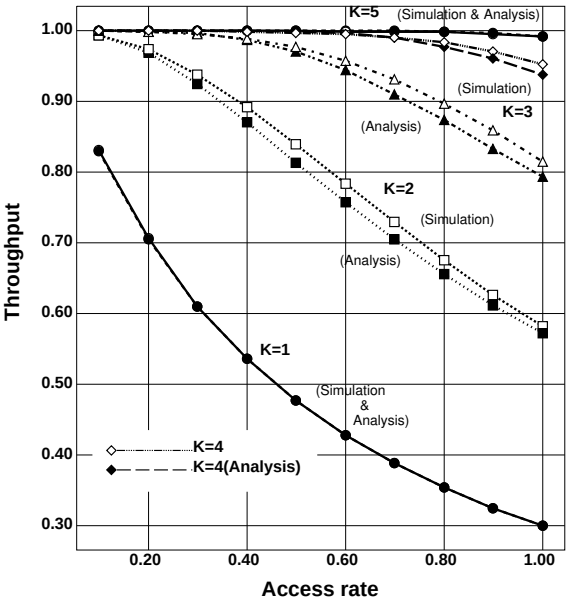

図 3: 負荷率と通過率の関係 (TBSF,$N = 256$)

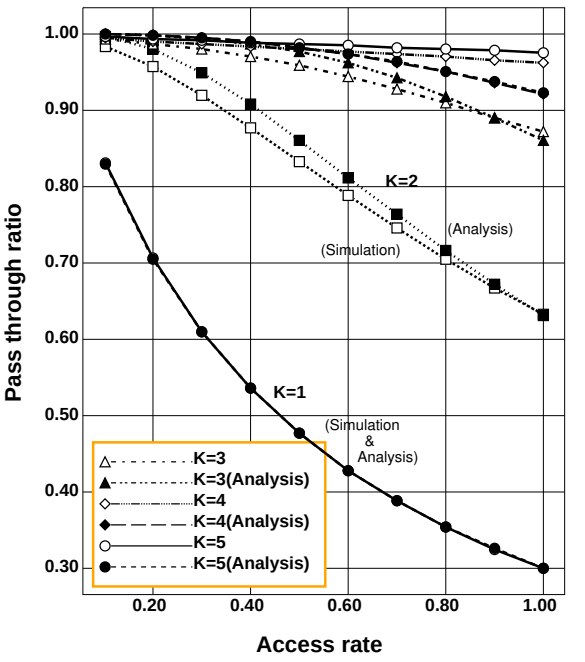

図 4: 負荷率と通過率の関係 (PBSF,$N = 256$)

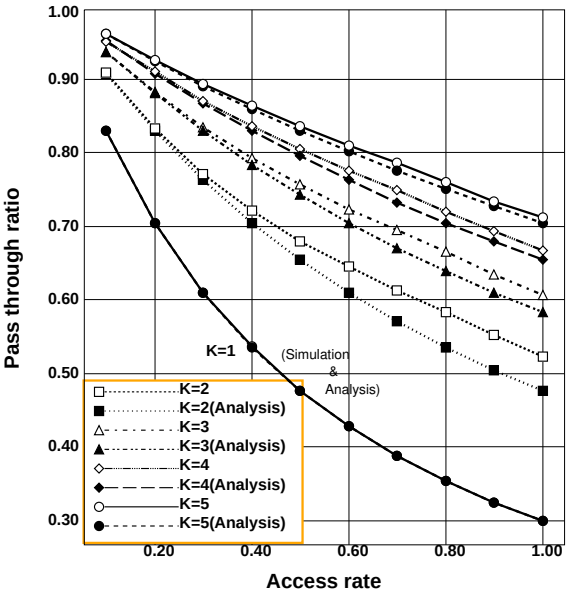

図 5: 負荷率と通過率の関係 (MBSF,$N = 256$)

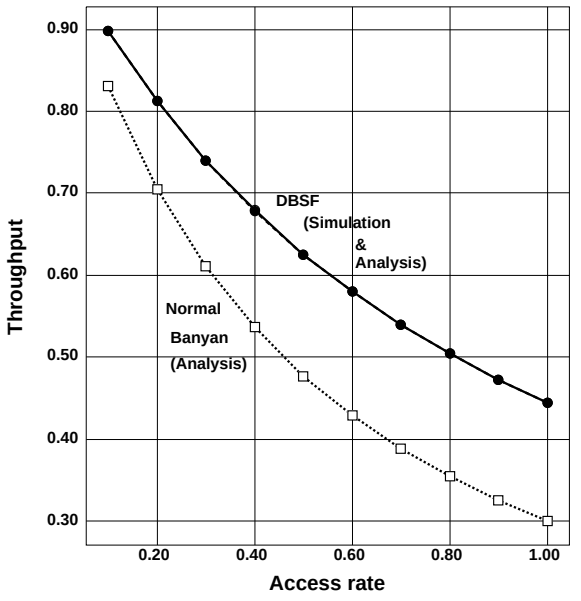

図 6: 負荷率と通過率の関係 (DBSF,N=256)

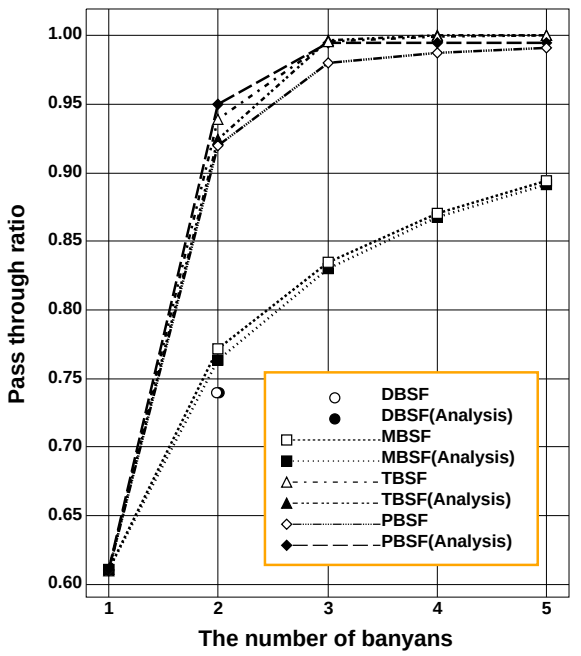

図 7: banyan 網数に対する通過率 (負荷率 30%,$N = 256$)

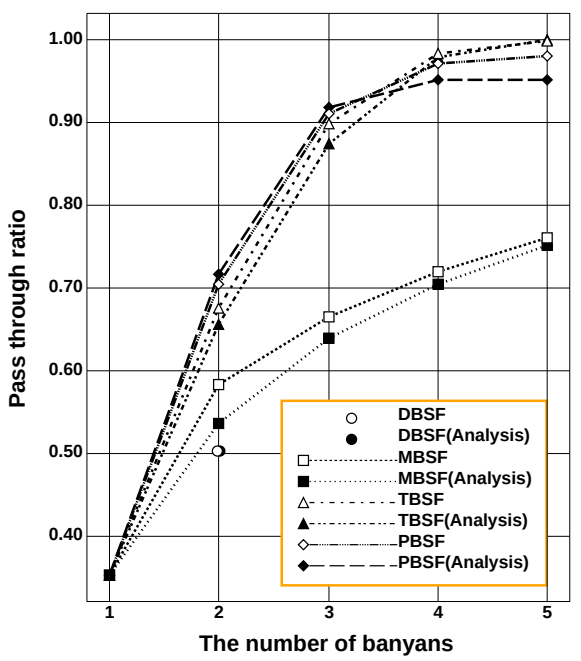

図 8: banyan 網数に対する通過率 (負荷率 80%,$N = 256$)

また、図 3,4 と 図 5,6 を比較すると、TBSF,PBSF では負荷率がある程度まで大きくなっても、通過率はわずかに減少するだけだが、MBSF,DBSF では負荷率が少しでも増加すると、通過率は急激に減少している.

これは、TBSF,PBSF のような網では、banyan 網を相互に接続し、競合に敗れたパケットを他の banyan 網で再ルーティングすることができるため、網の負荷が大きくなっても通過率はあまり大きく低下することはないからであると思われる.

さて、ここで TBSF と PBSF の得失について考えてみる. 図 3と 4、また図 7 ,8より、TBSF と PBSF の通過率にはほとんど差は見られない.

しかしながら、図 1と図 2を比べてわかる通り、TBSF では banyan 網の段数を増やすと、段数に比例してパケットが到着するまでの遅延が増加する. 一方、PBSF では 垂直方向のパスが存在するため、スイッチングエレメントが複雑になってしまい、通常のスイッチングエレメントと比較して 1.2-1.5 倍のハードウェア量が必要となる. しかし、現在の VLSI の実装技術であれば実現可能であると思われる.

従って、PBSF の方が TBSF より優れた性質を持つことがわかる.

6 データのアラインメント

本章では、データアクセスの高速化の一手法として、データのアラインメントについて検討する. Lawrie は非衝突アクセスのためのデータ ・アラインメントとして 2 倍のメモリ ・モジュール、DBSF を用いる方法を提案している [2]. 多重出力可能な MIN を用いることで、同一メモリ ・モジュールに対して複数のアクセスが可能となるので、Lawrie による方法の応用によって TBSF,PBSF における非衝突アクセスが実現できる.

6.1 仮定

$N = n^2$個のプロセッサ,$M = N$個のメモリユニットを考える ($n = 2^m$ m: 任意定数) .

ここで用いる MIN を次のように仮定する.

ネットワークトポロジ ：TBSF,PBSF で用いる banyan 網として, Inverse Omega Network を仮定する.

Banyan 網の段数 (K) ：2

行列の “行, 列, 対角要素, 逆対角要素” の各部分へのアクセスを考える.

Lawrie は $2N$ のサイズのメモリ ・モジュールを用いて $N \times N$-行列の要素 a_{jk} を $j \cdot (N^{\frac{1}{2}} + 1) + k \cdot 2$ 番目のメモリ ・モジュールにストアする方法を提案している [2].

ここでは,$N \times N$ の TBSF/PBSF を仮定しているので, 要素 a_{jk} がメモリの $\lfloor (j \cdot (N^{\frac{1}{2}} + 1) + k \cdot 2)/2 \rfloor$ にストアされるという具合に 2 個のメモリ ・モジュールを 1 個に圧縮する (図 9).

MEMORY

0	1	2	3
a_{00}	a_{01}	a_{02}	a_{03}
a_{13}	a_{10}	a_{11}	a_{12}
a_{21}	a_{22}	a_{23}	a_{20}
a_{30}	a_{31}	a_{32}	a_{33}

図 9: skewing の方法

TBSF/PBSF 内で, パケットはヘッダの情報に従って転送されるため, 非衝突のアクセスを実現するには, 衝突を起こすと予測されるパケットのプライオリティを制御する必要がある.

ここでは, 偶数入力端からのパケットには高いプライオリティを与え, 奇数入力端からのパケットには低いプライオリティを与えるものとする.

図 9に示した skewing 法を用いると, メモリ ・モジュール M からプロセッサ N への写像 $P=\{(x,y)|0\le x,y<N\}$ は次のようになる (b は定数) ：

$$\text{rows}\quad P=\{(x,x+b)|0\le x<N\} \tag{14}$$

$$\text{columns}\quad \begin{cases} P=\left\{\left(x,\frac{x}{2}(n+1)+b\right)\;\middle|\;0\le x<N\right\} \\ P=\left\{\left(x,\frac{x}{2}(n+1)-\frac{1}{2}+b\right)\;\middle|\;0\le x<N\right\} \end{cases} \tag{15}$$

$$\begin{matrix}\text{forward}\\ \text{diagonals}\end{matrix}\quad \begin{cases} P=\left\{\left(x,\frac{x}{2}(n+3)+b\right)\;\middle|\;0\le x<N\right\} \\ P=\left\{\left(x,\frac{x}{2}(n+3)-\frac{1}{2}+b\right)\;\middle|\;0\le x<N\right\} \end{cases} \tag{16}$$

$$\begin{matrix}\text{backward}\\ \text{diagonals}\end{matrix}\quad \begin{cases} P=\left\{\left(x,\frac{x}{2}(n-1)+b\right)\;\middle|\;0\le x<N\right\} \\ P=\left\{\left(x,\frac{x}{2}(n-1)-\frac{1}{2}+b\right)\;\middle|\;0\le x<N\right\} \end{cases} \tag{17}$$

(上段 $x=$ 偶数, 下段 $x=$ 奇数)

ここで,TBSF/PBSF を用いた非衝突アクセスを議論するために必要な, 次の定義と定理を示す.

定義 Ω_N を $N\times N$ の Omega 網とする ($N=2^l$). Ω_N が P_N によって表された写像を 生成することができるならば, $\Omega_N\uparrow P_N$ と呼ぶ.

定義 $\text{TBSF/PBSF}_{N,K}$ を $N\times N$,K 段の TBSF/PBSF とする ($N=2^l$).$\text{TBSF/PBSF}_{N,K}$ が P_N によって表された写像を生成することができるならば,$\text{TBSF/PBSF}_{N,K}\uparrow P_N$ と呼ぶ.

定理 **1** $(ax+b,cx+d)$ が 入力 $ax+b \pmod N$ から出力 $cx+d \pmod N$ への結合を表すとき,

$$P_N(a,b,c,d,\xi)=\{(ax+b,cx+d)|0\le x<\xi\} \tag{18}$$

が $\gcd(a,N)\ge\gcd(c,N)$ かつ $\xi\le N/\gcd(a,N)$ を満たすならば, $\Omega\uparrow P_N(a,b,c,d,\xi)$ である.

この定理は Lawrie によって証明されている [2].

ここでは, $\alpha=\gcd(a,N),\gamma=\gcd(c,N),a'=a/\alpha$ and $c'=c/\gamma$ と定義する.

6.2 PBSF における非衝突アクセスの実現

定理 **2** $PBSF_{N,2}\uparrow P(rows)$

証明 定理 1 より, α かつ $\gamma=1$ なので $\alpha\ge\gamma$ であり,$\xi=N/2$ なので $\xi\le N/\alpha$ である. 従って $PBSF_{N,2}\uparrow P(\text{ rows })$ である. **Q.E.D**

定理 **3** $PBSF_{N,2}\uparrow P(columns)$

証明 ここでは $x=2m$ (偶数入力) のとき, そして $x=2m+1$ (奇数入力) のときの 二つの部分に分けて考える.

$$\begin{aligned} &\text{偶数}\ P=\left\{\left(2m,m(n+1)+b\right)\;\middle|\;0\le m<\frac{N}{2}\right\} \\ &\text{奇数}\ P=\left\{\left(2m+1,m(n+1)+\frac{n}{2}+b\right)\;\middle|\;0\le m<\frac{N}{2}\right\} \end{aligned} \tag{19}$$

となる.

(I) 偶数 :偶数入力からのパケットのプライオリティは, 奇数入力からのものより高いため, 偶数入力からのパケットは奇数入力からのものとは無関係にルーティングされる.

$n+1$ は 2 と互いに素なので $\gamma=1$,$\alpha=2$ より$\alpha\ge\gamma$, かつ $\xi=N/2$ より$\xi\le N/\alpha$ なので条件は満たされる.

(II) 奇数 :奇数入力からのパケットのプライオリティは低いので, 競合に敗れると, 第二層に送られることになる. しかし, 第一層での衝突以前のルーティングは有効なので, パケットは単一のオメガ ・ネットワーク上でルーティングされるものと考えてよい.

偶数部分の場合と同様$\alpha=2,\gamma=1$ より $\alpha\ge\gamma$, また $\xi\le N/\alpha$ であり, 条件が満たされる.

(I) と (II) より, $\text{PBSF}_{N,2}\uparrow P(\text{ columns })$ である. **Q.E.D**

対角要素, 逆対角要素の証明は定理 3と同様である.

6.3 TBSF における非衝突アクセスの実現

定理 **4** c が奇数である時, 入力 $2n$ または $2n+1 \pmod N$ から 出力 $cm+d$ または $cm+d' \pmod N$ への結合は次のように表される ：

$$\begin{aligned} &\text{偶数}\quad P=\left\{(2m,cx+d)\;\middle|\;0\le m<\frac{N}{2}\right\} \\ &\text{奇数}\quad P=\left\{(2m+1,cx+d')\;\middle|\;0\le m<\frac{N}{2}\right\} \end{aligned} \tag{20}$$

全ての奇数入力からのパケットが一段目の $Banyan$ 網で出力に到達できないと仮定する. その時, これらのパケットは一段目の $Banyan$ 網の終端では次のような出力へルーティングされる ：

$$P(\text{miss})=\left\{\left(2m+1,(N-c)m+d\right)\;\middle|\;0\le m<\frac{N}{2}\right\}\quad \text{(奇数)} \tag{21}$$

証明 $N-c$ は 2 と互いに素なので,$\gamma=1$ を得, また $\alpha=2$ である. そこで $\alpha\ge\gamma$ であり, また $\xi=N/2$ なので $\xi\le N/\alpha$ である. 従って定理 1より TBSF は $P(\text{miss})$ を通過させることが分かる. **Q.E.D**

定理 **5** 式 20の条件を満足するとき, 一段目の $Banyan$ 網の終端において, 本来の出力にルーティングできない奇数入力からのパケットは, 二段目の $Banyan$ 網を通過する.

証明 一段目の Banyan 網の終端では, 式 4に示されるようにルーティングされる. 式 20に示される本来の出力にルーティングできない奇数入力からのパケットは, 二段目の Banyan 網を次のような規則で通過する.

$$\begin{aligned} &\text{TBSF}_{N,2}\uparrow P(\text{miss}\to\text{right}) \\ &=\left\{\left((N-c)m+d,cm+d'\right)\;\middle|\;0\le m<\frac{N}{2}\right\}\quad \text{(奇数)} \end{aligned} \tag{22}$$

一段目の Banyan 網の終端では, パケットは式 4のようにルーティングされる. さらに, これらのパケットは二段目の Banyan 網においては式 20に示される本来の出力へ転送される. c は奇数なので,$\gamma=1$,$N-c$ は 2 と互いに素なので, $\alpha=1$ である. こうして$\alpha\ge\gamma$, $\xi=N/2$ となり,$\xi\le N/\alpha$ である.

定理 1より, $\text{TBSF}_{N,2}\uparrow P(\text{ miss }\to\text{ right })$ である.

Q.E.D

定理 **6** $TBSF_{N,2}\uparrow P(rows)$

証明 PBSF の row の場合と同様である.

定理 **7** $TBSF_{N,2}\uparrow P(columns)$

証明 この証明では,$x=2m$ (偶数入力) と $x=2m+1$ (奇数入力) の二つの部分に分ける.

$$\begin{aligned} &\text{偶数}\ P=\left\{\left(2m,m(n+1)+b\right)\;\middle|\;0\le m<\frac{N}{2}\right\} \\ &\text{奇数}\ P=\left\{\left(2m+1,m(n+1)+\frac{n}{2}+b\right)\;\middle|\;0\le m<\frac{N}{2}\right\} \end{aligned} \tag{23}$$

偶数入力からのパケットのプライオリティは奇数入力よりも高いので, 一段目のネットワークを通過できる. 次に, 第二段のネットワークでは, 奇数部分が転送され, そこで全て通過できることを示す.

(I) 偶数 ：$n+1$ は 2 と互いに素なので $\gamma=1$, また $\alpha=2$ より $\alpha\ge\gamma$ であり, $\xi=N/2$ なので $\xi\le N/\alpha$ である. 従って定理 1より, 偶数の部分は一段目の Banyan 網を全て通過する.

(II) 奇数 ：パケットは定理 4のようにルーティングされるので,

$$\begin{aligned} P(\text{right}) &= \left\{\left(2m+1,m(n+1)+\frac{n}{2}+b\right)\;\middle|\;0\le m<\frac{N}{2}\right\} \\ &\Downarrow \\ P(\text{miss}) &= \left\{\left(2m+1,m\big(N-(n+1)\big)+N-(n+1)+b\right)\right. \\ &\qquad\left.\;\middle|\;0\le m<\frac{N}{2}\right\} \end{aligned} \tag{24}$$

となる.

次に, 二段目では,$n+1$ が 2 と互いに素なので,$\gamma=1$ である. そこで, 定理 5より, 奇数入力からのパケットは, 二段目の Banyan 網上で正しい行き先に送られる. 従って, $\text{TBSF}_{N,2}\uparrow P(\text{ columns })$ である.

Q.E.D

対角要素, 逆対角要素に対する証明は定理 7と同様である.

7　おわりに

多重出力を許す MIN はクロスバのようなブロッキング網よりも高い通過率を持つ. 中でも TBSF,PBSF は、MBSF,DBSF といった網に比べ、高い負荷率の下でも高い性能を示す. 実装をする際の banyan 網数は、2-3 段が現実的な値であるが、今回の解析、シミュレーションの結果、負荷率 100%という極端な状況を仮定しても 2 段の時には通過率 60%程度であり、3 段使用すると、90%近い通過率が得られることがわかった. 従って、TBSF,PBSF を用いることによって、従来の MIN と違ってバッファを持たない SSS 型 MIN でも十分な通過率を得ることができる. TBSF では、レイテンシィが大きく、その分性能が低下するおそれがあるが、PBSF では、一段の banyan 網のレイテンシィに加えて数クロック余計にかかるだけであり、レイテンシィの面からも優れた結合網である. ただし、入力数がある程度多くないと、メリットは少ない.

さらに、データのアラインメントを施すことによって、行列の各要素に対して非衝突にアクセスできることを示した. それによって、さらに高いパフォーマンスを得ることができる.

参考文献

[1] Clos,C. : "A study of non-blocking switching networks," Bell System Tech. J. 32, 2 Mar. (1953).

[2] Lawrie,D.H. : "Access and Alignment of Data in an Array Processor," IEEE Trans. on Comput. vol. c-24, No.12, Dec. (1975).

[3] Broomell,G. and Heath,J.R. : "Classification Categories and Historical Development of Circuit Switching Topologies," ACM Computing Surveys, Vol.15, No.2, Jun. (1983).

[4] Kruskal, C.P. and Snir, M. : "The performance of multistage interconnection networks for multiprocessors," IEEE Trans. Comput. Vol.C-32, No.12, pp.1091-1098, Dec. (1983).

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

[6] Kumar, M. and Jump, J.R. : "Performance of unbuffered shuffle-exchange networks," IEEE Trans. Comput. Vol.C-35, No.6, pp.573-577, Jun. (1986).

[7] 坂元, 荒川, 正木, 井上, 天野: 自己ルーチングスイッチの構成とその評価, 信学技報 ISSE88-30 8, (1988).

[8] Lea, C.T. : "Multi-Log_2N networks and their applications in high-speed electronic and photonic swtiching systems," IEEE Trans. Comm. 38(10) pp.1740-1749, (1990).

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

[10] Tobagi, F.A. Kwok, T.:*The Tandem Banyan Switching Fabric: a Simple High-Performance Fast Packet Switch*, Proc. of INFOCOM91, (1991).

[11] Chiussi, F. Amano, H. Tobagi, F.A.: *A 0.8μm BiCMOS Sea-Of-Gates implementation of the Tandem Banyan Fast Packet Switch*, Proc. of IEEE Custom Integrated Circuits Conference, (1991).

[12] Amano, H. Zhou, L. Gaye, K.:*SSS(Simple Serial Synchoronized)-MIN: A novel multi stage interconnection architecture for multiprocessors*, Proc. of IFIP Congress 92, Sept. pp. 571-577, (1992).

[13] Wu, C.L. Lee, M. : "Performance Analysis of Multistage Interconnection Network Configurations and Operations", IEEE. Trans. Comput. **41**, No.1 pp.18-27, (1992).

[14] Gaye, K. Amano, H.:*A Batcher Double Omega network with Combining*, IEICE Trans. on information and systems, Vol.E75-D No.3 (1992).

[15] 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).

[16] 天野英晴, 藤川義文:マルチステージネットワーク PBSF, 情報処理学会計算機アーキテクチャ研究報告 No.94-5, (1992)

[17] 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).