多重出力可能な MIN の性能評価*

塙 敏博　　天野 英晴†
慶應義塾大学 理工学部‡

概要

多重出力可能な MIN(Multistage Interconnection Network) は、同一宛先に対して複数パケットの通過を可能にすることにより、高いスループットを実現する多段結合網である。その高いスループットにより、SSS(Simple Serial Synchronized) 型 MIN として並列計算機に用いることができると共に、ATM(Asynchronous Transfer Mode) 用パケット交換機にも利用することができる。

本論文では、複数の Banyan 網にトラフィックを分散する MBSF(Multi-Banyan Switching Fabrics)、サイズの大きい Banyan 網を用いる EBSF(Expanded Banyan Switching Fabrics)、Banyan 網を縦列接続する TBSF(Tandem Banyan Switching Fabrics)、三次元的に接続する PBSF(Piled Banyan Switching Fabrics) の 4 種類の多重出力可能な MIN についてその通過率を理論モデルとシミュレーションによって解析した。その結果、接続 Banyan 網数が少ない領域では PBSF が有利で、多い領域では TBSF が有利であることを示した。

さらに、並列計算機に用いた場合、これら全ての網について、メモリモジュールへのデータのアラインメントを適切に行なえば、行列の要素 (行・列・対角要素・逆対角要素) への非衝突アクセスが可能であることを示した。

Abstract

Multistage Interconnection Networks(MINs) with multiple outlets are networks which can support higher bandwidth than those of nonblocking networks by passing multiple packets to the same destination. They are not only useful as a MIN topology for the Simple Serial Synchronized (SSS)-MIN used in multiprocessors, but also as a switching core in the ATM(Asynchoronous Transfer Mode) packet exchanger.

Four MINs with multiple outlets: Multi-Banyan Switching Fabrics (MBSF), Expanded Banyan Switching Fabrics (EBSF), Tandem Banyan Switching Fabrics (TBSF) and Piled Banyan Switching Fabrics (PBSF) are analyzed and compared with both the theoretical model and simulation.

The PBSF supports the best throughput and latency used for the SSS-MIN. Although the latency of the TBSF is large, the pass-through ratio is the best if the number of connected banyan networks are more than 4. Therefore, the TBSF is useful for the ATM switching networks in which the relatively large latency is tolerable.

The conflict-free access of these MINs is also analyzed, and it appears that rows, column, forward and backward diagonal of the matrix can be accessed without conflict.

1　はじめに

多段結合網 (Multistage Interconnection Network: MIN)[3] は中規模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．従来の代表的な MIN では, メモリに対するアクセスは, パケットの形で行われ, スイッチングエレメント間はビットパラレル (8 ~ 64bit) で転送される．また, トポロジとしては omega 網に代表されるブロッキング網が使用され, スイッチングエレメント内で衝突が起きた場合, 片方のパケットは内部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, 高速化, 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, このような従来型の MIN ではプロセッサ当たり 8 ~ 64bit の入出力線が必要なため,LSI チップ中のエレメント数がピン数によって制限される．このことは, 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を困難にし, エレメント構造と制御が複雑になる．そして 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,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 [17].

SSS 型 MIN は, 高密度実装と高い周波数のクロックの利用が可能である一方, 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は再送を行なう必要があり, 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 で用いられてい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なため, 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, 63.5%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], 整数倍のサイズの Banyan 網を用いる Expanded-Banyan[18](Double-Banyan[14]) が知られている．本論文では, これらの MIN の通過率を確率モデルに基づいて解析し, 比較検討を行う．また, 科学技術計算等で頻繁に使用される行列のアクセスの能力に関しても理論的に検討し, 比較する．

---

*Performance analysis of MINs with multiple outlets

†Toshihiro Hanawa　Hideharu Amano

‡Keio University

2 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.

2.1 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) パケット交換用に, 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.

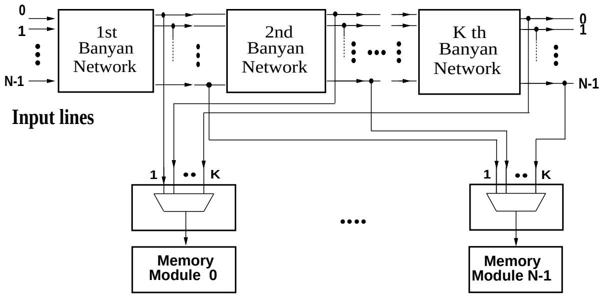

図 1: TBSF の構成

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.banyan 網を通過して目的の宛先に到着したパケットはバイパス路によりメモリモジュールに送られ, 衝突により目的の宛先に到着できなかったパケットのみが次の段の banyan 網に入力される. 衝突の際にセットされた conflict bit は, 次の網の入力時にリセットされる.

TBSF を B-ISDN 用の交換機に用いる場合には, パケットは各 banyan 網から時間的にずれて到着するので, それ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, パケットの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 フレーム内に何回のメモリアクセスが可能であるかは, 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と同様に 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].

2.2 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. 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下の層のネットワークに送られる.

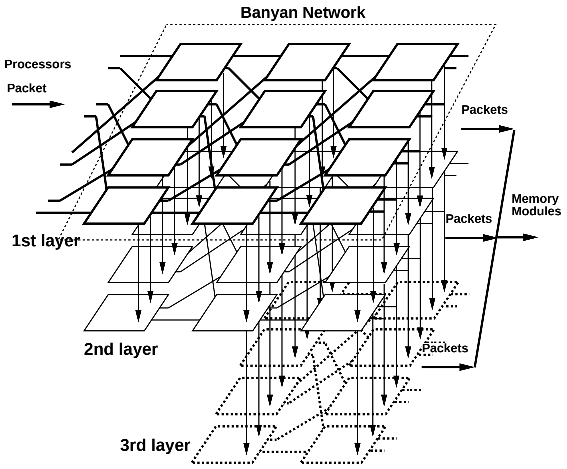

図 2: PBSF の構成

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.

水平方向に進んでくるパケットと上層から送られたパケットとが競合すると, 上層からのパケットが優先的に水平方向に送られ, 水平方向に進んできたパケットはさらに下の階層に送られる.

水平方向からのパケット 2 つと, 上層からのパケット (1 つしか存在し得ない) がすべて同一出力線を目指した場合には, 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一方が下層に送られる. 残った 1 つのパケットは衝突したことを示す conflict bit がセットされ, 希望しない方の水平方向の出力に送られる.

最下層のネットワークで水平方向からのパケット同士が競合に敗れた場合も同様に conflict bit がセットされ, 以後他のパケットを妨害しない. 最下層において水平方向から 2 つ, 垂直方向からのパケットが競合した場合には垂直方向からのパケットは正常に送られ、水平方向からのパケットのうち一方は conflict bit がセットされ、もう一方のパケットは出力先がないためエレメント内で消滅する。

PBSF において下層の網は, 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. 最上層以外の網の出力からは希望の出力に到着できなかったパケットが出力される可能性があり, TBSF 網同様,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.

PBSF のスイッチングエレメントは垂直方向の線が, 各エレメントの出力線に対して用意されている. 上層での競合に敗れたパケットは希望の出力に対応する垂直方向の線を

通って半クロック分の遅延で下の層に送られ, その層での水平方向の出力線を確保する. ここで, さらに半クロック (上層に比べると 1 クロック) 遅れて水平方向にパケットが到着し, コントローラ部で進路が判定される. 目的の出力線がすでに上層からのパケットで占拠されていた場合, 垂直方向の線を通って下層に送られる. すなわちこの方式は垂直方向優先である.

2.3 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 TBSF, PBSF にならって,MBSF(Multi Banyan Switching Fabrics) と呼ぶ. TBSF,PBSF においては各 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だけである.

2.4 EBSF(Expanded Banyan Switching Fabrics)

Banyan 網のサイズを整数倍に拡張すると、通過率に関して MBSF と同等程度の能力を持つものと考えられる [18]. 他の網における Banyan 網数と EBSF の Banyan 網のサイズが対応する. 2 倍のものについては Lawrie によって行列に対するアクセス能力が解析されている [2]. 他の網にならって EBSF(Expanded Banyan Switching Fabrics) と呼ぶ.

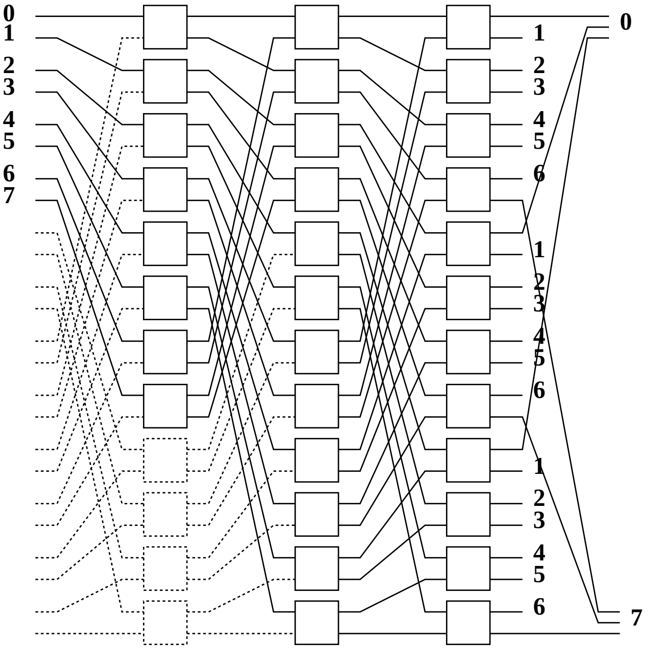

図 3: EBSF の構成

図 3は 入力数 8, banyan 網のサイズ 3 倍の EBSF を示している. 入力が 8 なので最初のステージの 12 のエレメントのうち 8 つだけが使われ, 残りの 4 つのエレメントは使用されない.

2.5 定性的な評価

TBSF,PBSF,MBSF,EBSF 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 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ない. EBSF では MBSF と違って 1 つの banyan 網であるため MBSF よりは通過率は高いが入力数よりもステージあたりのエレメント数が多い場合には, 前半のステージでは使われないエレメントが生じるので無駄が大きい. PBSF は, 競合に敗れたパケットも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 縦列接続を行う TBSF が最も不利である. 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.EBSF,MBSF は通常の Banyan 網と同じで通過時間は log_2Nとなる. これに対し,TBSF の通過時間は $K \times log_2N$,PBSF は $K-1+log_2N$である. すなわち,K が小さい時, レイテンシィは MBSF, EBSF, PBSF は,TBSF に比べて,$1/K$程度になる.

コスト エレメント数は 4 種類すべてについて $(KN)/2log_2N$ である (EBSF, PBSF の下層の網の前の方のいくつかは省略できる). エレメントの構造については PBSF 以外のスイッチングエレメントは 2 入力 2 出力であるが PBSF の場合は垂直方向の入出力が加わり,4 入力 4 出力となる. このエレメントは, 水平方向の出力に対して垂直方向の入出力線が設けられるためハードウェア量は 4×4 のクロスバよりは小さいが, それでも他の網のエレメントの 1.5 ~ 1.8 倍程度のハードウェアを必要とする.

3 解析モデル

均一なトラフィックを仮定した場合の SSS 型 MIN の通過率を解析するための確率モデルを構築する. このモデルは, Wu らによるサーキットスイッチングを行なう MIN の解析手法 [13] に似ているが, 多重出力である点, パケットが同期入力される点が異なる.

解析は以下の仮定に基づく.

- バッファは持たない
- 結合網に対する各入力へのパケットの入力確率は均一である
- パケットの行き先のアドレスは等確率に分散される
- パケットは全て同期して入力される
- 目的のアドレスに到着しなかったパケットはそのまま捨てられる
- 各スイッチングエレメントは PBSF では 4 入力 4 出力, それ以外では 2 入力 2 出力を持つ

3.1 MBSF

まず, 各スイッチングエレメントの状態について考える. ここでは スイッチングエレメントにパケットが入力される確率を r と仮定する. また, パケットが両方の入力から到着し同じ出力に向かうときに衝突が起こるので パケットの衝突確率は $0.25r^2$ となる. SSS 型 MIN では衝突によって正しくない出力へ送られたパケットはデッドパケットとなり, 他のパケットを妨害しない, すなわちパケットは消滅したものとみなされる.

i 番目のステージのスイッチングエレメントにパケットが入力される確率を r_i とすると 次のステージ $i+1$ での入力確率 r_{i+1} は以下の通りである.

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

従って 1 段,n ステージを持つ banyan 網の通過率 (パケットが正しい出力に到着する確率) は次のように表される

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

ここで r_0 は MIN 全体に対するパケットの入力確率を表す.

MBSF では 入力されたパケットは均等に K 層の banyan 網に分散される. 従って MBSF の通過率 PT_{MBSF} は以下のようになる.

$$PT_{MBSF} = KX^n(r_0/K) \tag{3}$$

3.2 TBSF

TBSF では 正しくルーティングされたパケットはメモリモジュールに送られ, 次の banyan 網ではそのパケットが取り除かれる. 従って,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 とすると, 次の式が成立する.

$$R_k = R_{k-1} - X^n(R_{k-1}) \tag{4}$$

この漸化式から TBSF の通過率 PT_{TBSF} は以下のように表される.

$$PT_{TBSF} = R_0 - R_K \tag{5}$$

ここで K は banyan 網数である.

3.3 PBSF

PBSF では以下の仮定に基づく.

- 上層から垂直方向に送られるパケットは水平方向からのものより優先される

PBSF の最上層では パケットが水平方向に送られる確率は式 1で表される. PBSF では衝突が起こるとそのパケットは下層に送られる. 従って パケットが第 2 層の i 番目のステージに垂直方向から入力される確率 $d_{i,2}$ は次のように表すことができる.

$$d_{(i,2)} = 0.25{r_i}^2 \tag{6}$$

二段目以降の banyan 網について, 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.

この表を用いて各スイッチングエレメントについて水平方向と垂直方向の出力確率を求め, 順次計算を行う.

3.4 EBSF

最初のステージでは EBSF の入力 N に対し, エレメントの入力の総数は cN （c は網のサイズを表す定数）であるので,cN 個のエレメントのうち N 個のエレメントの入力の片方のみにパケットが入力され,$(c-2)N/2$ 個のエレメントには入力されない (すなわち そのエレメントは必要ない). 従って パケットが出力される確率は N 個のエレメントについては $r_0/2$, $(c-2)N/2$ 個のエレメントでは 0 である.

表 1: スイッチングエレメントのパケットの出力確率 (PBSF)

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

サイズが奇数倍の場合は, ステージ 2 以降でも 等しい入力確率を持つエレメントの組が 2 つでき, その個数は N と $(c-2)N/2$ または $2N$ と $(c-4)N/2$ である. エレメントの入力確率をそれぞれ p,q とすると パケットが通過する確率は $(p+q)/2 - pq/4$ となる.

サイズが偶数倍の場合には 同じステージのエレメントは全て等しい入力確率を持つ. 従ってパケットが通過する確率は 1式と同様である.

4 シミュレーションと評価

図 7,8,9,10はそれぞれ EBSF,MBSF,TBSF,PBSF の入力負荷率 (r_0) に対する通過率を示している. ここでは 網のサイズは 256 と固定し, 接続する banyan 網数 (K) を変化させている. また 確率モデルについて検証するために スイッチングエレメントのハードウェアを基にしたシミュレーションによる結果も示した.

解析モデルの仮定で, 入力負荷として 均一なトラフィックを仮定している. しかし TBSF と PBSF では 前の banyan 網によって トラフィックの均一さが失われるので, banyan 網に送られるトラフィックは均一ではない. このことが 解析モデルとシミュレーションとの間での わずかなずれとなって現れている. しかしながら この差は高々5%程度であり, 解析モデルが大部分の範囲で適当であったことを示している.

これらの図から, EBSF と MBSF では TBSF と PBSF に比べて 網を多重化したことによる通過率の改善はあまり大きくないことがわかる. これは EBSF ではサイズを大きくしたことによる効果は後半のエレメントにしか反映されないこと,MBSF では各 banyan 網が独立に動作するためだと考えられる. banyan 網 5 段の TSBF と PBSF では, 負荷率を 100%にしても大部分のパケットは通過することがわかる.

図 11 は 入力負荷を 100%で固定した時の 接続 banyan 網数に対する通過率を示したものである. この図で水平に引かれた線はクロスバのようなノンブロッキング網の通過率 (63.5%) を示している. TBSF と PBSF はわずか banyan 網 2 段でノンブロッキング網よりも高い通過率を示す. EBSF,MBSF においても banyan 網 3 段でノンブロッキング網の性能を上回っている. このことは 複数の出力を持つことが有利であることを示す.

図 9と図 10から, 接続された banyan 網が 2,3 と小さいときには PBSF の通過率は PBSF よりも 7%ほど上回っている. PBSF では 衝突したパケットは次の層の同じ場所にあるエレメントに入力されるので衝突以前のルーティングが有効になる. 一方 TBSF では最初のステージからルーティングし直さなければならない.

しかしながら PBSF では スイッチングエレメント内で 3 つのパケットが 1 つの出力に集中するとそのうちのパケット 1 つは水平の正しくない方向へ送られ, そのまま目的の出力へは到達できない. そのため PBSF では banyan 網の段数が大きいときには 98%程度で飽和するが TBSF では接続 banyan 網数を大きくすれば 99.9%以上の通過率を得ることができる.

これらの結果から PBSF は 小さな遅延で大きな通過率を持つことから SSS-MIN のようなマルチプロセッサに適し, TBSF は 比較的大きな遅延を許し極端に高い通過率を必要とする ATM 交換機に適しているといえる.

5 多重出力可能な MIN を用いた非衝突アクセスの実現

科学技術計算で多用されるベクトルや行列演算には メモリモジュールへ非衝突でアクセスできることが重要になる. EBSF,MBSF,TBSF,PBSF のような MIN は 多重出力を持っているため, ブロッキング網に比べて高い能力を持つ. この能力を最大限活用して 適切なデータのアラインメントを行うことによって, ベクトル, 行列の様々な部分に対して非衝突のアクセスが可能であると考えられる.

Lawrie は 2 倍のサイズの omega 網と 2 倍のサイズのメモリモジュールを用いて非衝突アクセスを行うためのデータのアラインメントを提案している [2]. しかし EBSF,MBSF,TBSF,PBSF は 複数の出力 (2 つで良い) を持ち 2 倍のサイズのオメガ網と同等以上の性能を持つので, それらの網を用いた SSS 型 MIN では 同様のデータアラインメントを施すことによって非衝突のアクセスが可能になると考えられる.

5.1 仮定

まず $N = n^2$ を プロセッサの数、$M = N$をメモリモジュールの数と仮定する.(n は 2 のべき乗)

本論文では 以下のような MIN を仮定する.

- ネットワークトポロジ : EBSF,MBSF,TBSF,PBSF で用いる banyan 網として, 逆 omega 網 [2] を仮定する. (図 4)

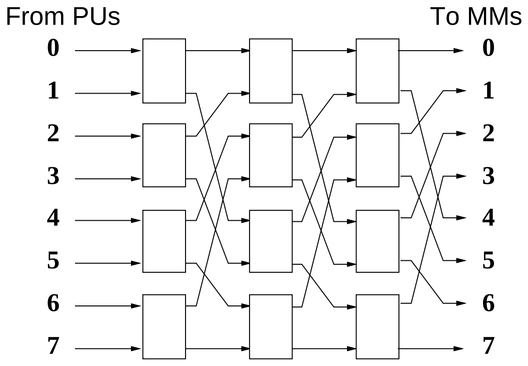

図 4: 逆 omega 網

- 使用する banyan 網の数 (K) : 2

ここでは 行, 列, 対角要素, 逆対角要素の 行列の一部への非衝突アクセスについて考える.

まず始めに メモリモジュール上のデータに対してアクセスの際に衝突を起こさないような データのアラインメントについて考える必要がある. データを skewing する方法が Lawrie によって提案されている [2]. この方法では $N \times N$行列の a_{jk} 要素は $j \cdot (N^{\frac{1}{2}} + 1) + k \cdot 2$ 番目のメモリモジュールにストアされる.

ここでは N個の入力を持つ EBSF/MBSF/TBSF/PBSF を仮定しているので $\lfloor (j \cdot (N^{\frac{1}{2}} + 1) + k \cdot 2)/2 \rfloor$ (図 6) のように 2 つのメモリモジュールが 1 つに圧縮される. この際, メモリモジュールへのアクセスは 1 フレームで 2 回のアクセスを許すなどが考えられる.

MEMORY

0	1	2	3	4	5	6	7
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 5: Lawrie による skewing 法

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 6: 本論文での skewing 法

EBSF,MBSF は [2] に示された 2 倍のサイズの omega 網と酷似しているので, 本論文では TBSF と PBSF についてのみ議論することにする.

TBSF/PBSF では パケットはヘッダに従って各スイッチングエレメントで独立にルーティングされる. 非衝突アクセスを実現するためには 衝突すると予測されるパケットのプライオリティを制御しなければならない. プライオリティビットはパケットのヘッダに加えられ, 高いプライオリティを持つパケットは低いプライオリティのものに優先する. ここでは偶数の入力からのパケットには高いプライオリティを与え、奇数の入力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 6に示した skewing 法を用いるときメモリモジュール 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).

$$\text{列:}\ P = \{(x, x + b) | 0 \le x < N\} \tag{7}$$

$$\text{行:}\ \left[\begin{array}{ll} P = \left\{ \left(x, \frac{x}{2}(n+1) + b\right) \middle| \ 0 \le x < N \right\} & (x = \text{偶数}) \\ P = \left\{ \left(x, \frac{x}{2}(n+1) - \frac{1}{2} + b\right) \middle| \ 0 \le x < N \right\} & (x = \text{奇数}) \end{array}\right. \tag{8}$$

$$\text{対角:}\ \left[\begin{array}{ll} P = \left\{ \left(x, \frac{x}{2}(n+3) + b\right) \middle| \ 0 \le x < N \right\} & (x = \text{偶数}) \\ P = \left\{ \left(x, \frac{x}{2}(n+3) - \frac{1}{2} + b\right) \middle| \ 0 \le x < N \right\} & (x = \text{奇数}) \end{array}\right. \tag{9}$$

$$\text{逆対角:}\ \left[\begin{array}{ll} P = \left\{ \left(x, \frac{x}{2}(n-1) + b\right) \middle| \ 0 \le x < N \right\} & (x = \text{偶数}) \\ P = \left\{ \left(x, \frac{x}{2}(n-1) - \frac{1}{2} + b\right) \middle| \ 0 \le x < N \right\} & (x = \text{奇数}) \end{array}\right. \tag{10}$$

TBSF/PBSF を用いた非衝突アクセスを議論するために必要な 以下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,

$$P_N(a,b,c,d,\xi) = \{(ax+b, cx+d) | 0 \leq x < \xi\} \tag{11}$$

が $\gcd(a,N) \geq \gcd(c,N)$ かつ $\xi \leq N/\gcd(a,N)$ 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.

5.2 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 $\alpha \geq \gamma$ であり,$\xi = N/2$ なので $\xi \leq N/\alpha$ 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.

$$\begin{aligned} &\text{偶数 } P = \left\{ \Big(2m, m(n+1)+b\Big) \;\middle|\; 0 \leq m < \frac{N}{2} \right\} \\ &\text{奇数 } P = \left\{ \left(2m+1, m(n+1)+\frac{n}{2}+b\right) \;\middle|\; 0 \leq m < \frac{N}{2} \right\} \end{aligned} \tag{12}$$

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$\alpha \geq \gamma$, かつ $\xi = N/2$ より$\xi \leq N/\alpha$ 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 $\alpha \geq \gamma$, また $\xi \leq N/\alpha$ 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.

5.3 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 :

$$\begin{aligned} &\text{偶数} \quad P = \left\{ (2m, cx+d) \;\middle|\; 0 \leq m < \frac{N}{2} \right\} \\ &\text{奇数} \quad P = \left\{ (2m+1, cx+d') \;\middle|\; 0 \leq m < \frac{N}{2} \right\} \end{aligned} \tag{13}$$

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 :

$$P(\text{miss}) = \left\{ \Big(2m+1, (N-c)m+d\Big) \;\middle|\; 0 \leq m < \frac{N}{2} \right\} \quad \text{(奇数)} \tag{14}$$

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 $\alpha \geq \gamma$ であり, また $\xi = N/2$ なので $\xi \leq N/\alpha$ 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 13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 13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.

$$\begin{aligned} &\text{TBSF}_{N,2} \uparrow P(\text{miss} \rightarrow \text{right}) \\ &= \left\{ \Big((N-c)m+d, cm+d'\Big) \;\middle|\; 0 \leq m < \frac{N}{2} \right\} \quad \text{(奇数)} \end{aligned} \tag{15}$$

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 13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$\alpha \geq \gamma$, $\xi = N/2$ となり,$\xi \leq N/\alpha$ である.

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

$$\begin{aligned} &\text{偶数 } P = \left\{ \Big(2m, m(n+1)+b\Big) \;\middle|\; 0 \leq m < \frac{N}{2} \right\} \\ &\text{奇数 } P = \left\{ \left(2m+1, m(n+1)+\frac{n}{2}+b\right) \;\middle|\; 0 \leq m < \frac{N}{2} \right\} \end{aligned} \tag{16}$$

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 $\alpha \geq \gamma$ であり, $\xi = N/2$ なので $\xi \leq N/\alpha$ 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,

$$\begin{aligned} P(\text{right}) &= \left\{ \left(2m+1, \underset{\Downarrow}{m}(n+1)+\frac{n}{2}+b\right) \;\middle|\; 0 \leq m < \frac{N}{2} \right\} \\ P(\text{miss}) &= \Big\{ \Big(2m+1, m\big(N-(n+1)\big)+N-(n+1)+b\Big) \\ &\qquad\qquad \left|\; 0 \leq m < \frac{N}{2} \right\} \end{aligned} \tag{17}$$

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.

6 おわりに

EBSF,MBSF,TBSF,PBSF という多重出力可能な MIN について解析し, 比較を行った. 複数の出力を持つため 同一宛先に向かう複数のパケットが通過できることでノンブロッキング網より高い通過率を得ることができる. PBSF は評価によって スループットとレイテンシの点で SSS 型 MIN に最適な網であることが示された. TBSF のレイテンシは大きいが,5 段以上の banyan 網を接続すれば通過率は最大になる. 従って TBSF は比較的大きなレイテンシでもあまり問題とならない ATM 交換機用の網に向いている.

本論文では 均一なトラフィックを仮定して解析を行ったが, 非均一なトラフィックのもとの解析については 参考文献 [19] を参照されたい. 但し TBSF のみについての試みであり, 他の網についての検討はこれからの課題である.

さらに, 行列の部分 (行, 列, 対角要素, 逆対角要素) に対する非衝突のアクセスが可能であることを示した.

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 英晴, 周 洛, 藤川 義文: 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).

[18] Awdeh, R. and Mouftah, H.: “The Expanded Delta Fast PAcket Switch”, IEEE Int. Conf. Commun. (ICC), (1994).

[19] Gaye, K., 塙 敏博, 天野 英晴: “SSS 型 MIN における hot spot の影響とメッセージ結合の効果,” 電子情報通信学会論文誌 D-I, (1994 年 5 月号掲載予定).

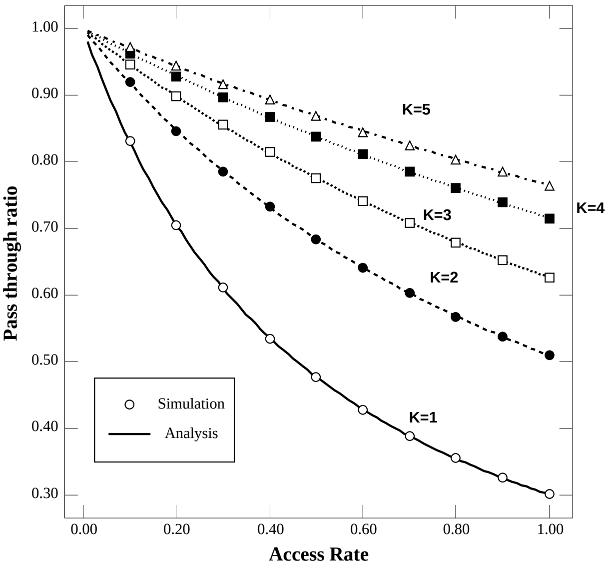

図 7: 負荷率と通過率の関係 (EBSF,256 入力)

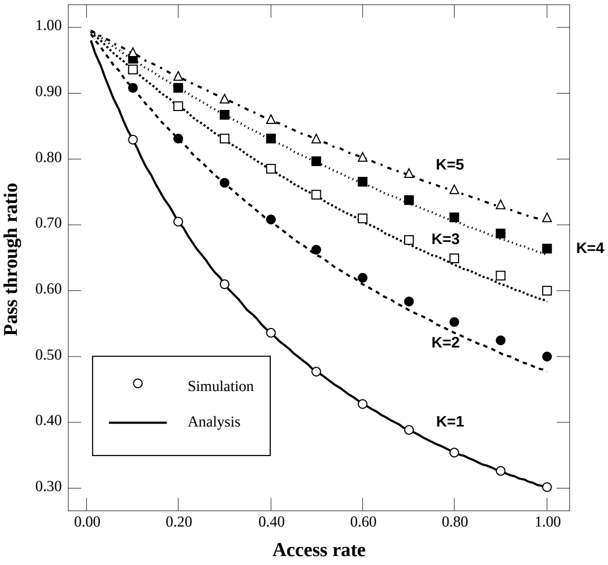

図 8: 負荷率と通過率の関係 (MBSF,256 入力)

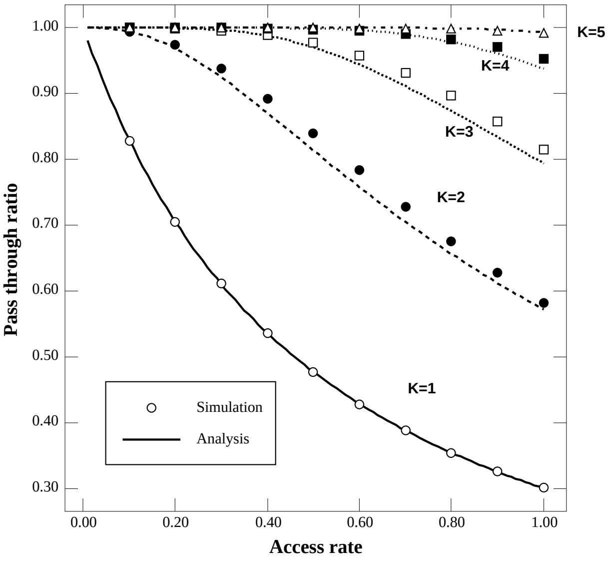

図 9: 負荷率と通過率の関係 (TBSF,256 入力)

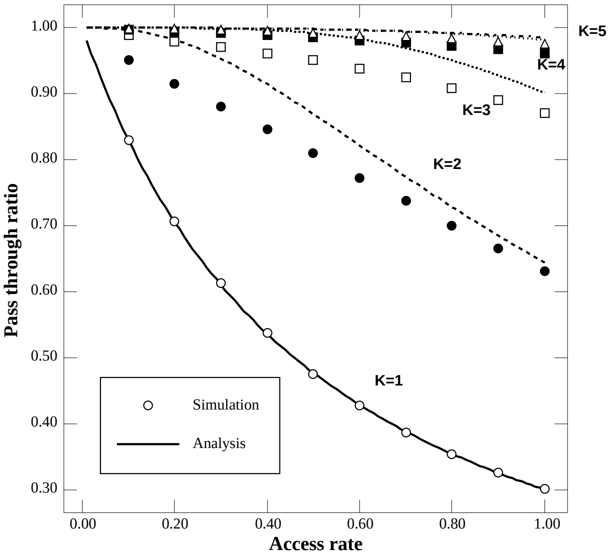

図 10: 負荷率と通過率の関係 (PBSF,256 入力)

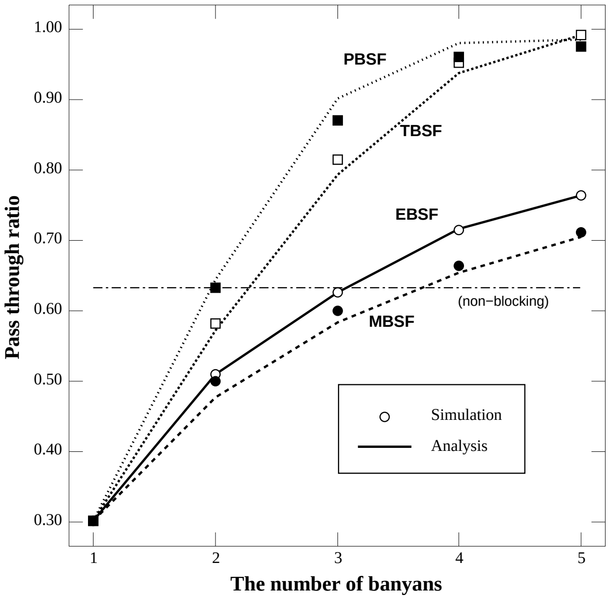

図 11: banyan 網数に対する通過率 (負荷率 100%)