1 はじめに

多重出力可能なMINは、同一宛先に対して複数のパケットの通過を 可能にすることにより、ノンブロキングネットワークに比べ高いスループットを実現する多段結合網である。

単純な多重出力可能な MIN として Multi Banyan Switching Fabrics (MBSF) [6][7]、Expanded Banyan Switching Fabrics (EBSF) [2] が提案されているが、理論解析の結果より [2]、これら2つのネットワークでは構成が単純すぎ通過率は低いことがわかっている。

より高性能な多重出力可能な MIN、 Tandem Banyan Switching Fabrics (TBSF) [4][3], Piled Banyan Switching Fabrics (PBSF) [10] は、MBSF,EBSF よりも高い通過率を示し [10]、マル チプロセッサでのプロセッサーメモリ間の多段結合網や ATM パケット交換網用と して有効である。

これら2つのネットワークは複数の MIN から構成されるため、潜在的に耐故障性を持っている。本論文では、故障回復メカニズムを上記2つのネットワークに付加した耐故障 TBSF(F-TBSF)、耐故障 PBSF(F-PBSF) を提案する。さらに、F-TBSF,F-PBSF のエレメントが故障したときの性能低下を確率モデルとシミュレーションによって解析した。

2 コントロールモデル、故障モデル

2.1 多重出力可能な MIN のコントロールモデル

多重出力可能な MIN は単純な構造とコントロールにより構成されるスイッチングシステムである。全てのパケットはフレームクロックに同期して入力パケットバッファからシリアル (数ビットシリアル) に入れられる。各スイッチングエレメントはパケットの 1 ビット (もしくは数ビット) だけをストアし、MIN はスイッチング能力を持つシフトレジスタのように動作する。

スイッチングエレメントの中にパケットバッファを持たないため、パケットの衝突が起きると、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。このとき、この希望の方向に進めなかったパケットの routing tag 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デッドパケットとして扱われ、他のパケットの進行を妨害しない。この単純なコントロール／構造のため、全てのスイッチングエレメントにパケットバッファを内蔵した MIN に比べて高い周波数で動作でき、高密度実装が可能である。これらの特徴により、ATM パケット交換網やマルチプロセッサ [1][11] で利用されている。

しかし、MIN 内でパケット同志の衝突が起きた場合は衝突したパケットを次のフレームで再送を行う必要があり、性能低下の大きな原因となる。この場合に、多重出力可能な MIN は同一宛先に対して複数のパケットの通過が可能なため、有利である。

2.2 故障モデル

一般の MIN と同様に、多重出力可能な MIN は単純な 2×2 や 4×4 のスイッチングエレメントにより構成される。図 1 で示すように、パスに付加されたコントローラがパケットのヘッダをチェックし、マルチプレクサをセットすることによりスイッチングエレメントを' straight' か 'cross' にする。ここで、ほとんどの MIN の故障モデル [9],[12] と同様に以下の 2 つのタイプの故障について考える。

1. マルチプレクサ内、またはスイッチングユニット間のリンクの故障 (link fault)

2. コントローラとマルチプレクサの誤動作 (element fault)

前者の故障は故障したリンクに routing されたパケットの消失を起こし、後者の故障の方はスイッチの 'stuck' を引き起こし、パケットのミスルーチングの原因になる。一般的に、コントローラの面積はスイッチングエレメント内の他の部品よりも大きく、'element fault' の方が 'link fault' よりも起きる確率が高い。ここで扱われるコントロールモデルは、パケットはシリアルに転送される。よって、パケットの部分的な損失 [13] は扱わない。

3 多重出力可能な耐故障 MIN

3.1 TBSF(Tandem Banyan Switching Fabrics)

TBSF は, 本来 B-ISDN(Broadband Integrated System Digital Network) 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 [4] に, 海外では Tobagi らにより 1990 年 [5] に, 提案された網である。

TBSF は, 図 2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。メモリモジュールへのインターフェースは各 banyan 網の出力のためにパケットバッファを持つ。

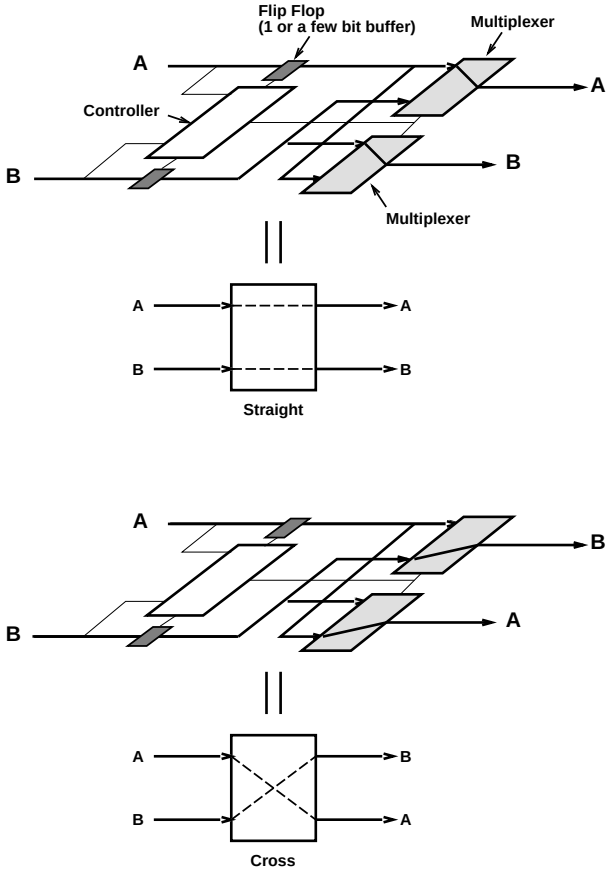

図 1: Switching element

3.2 TBSF の故障回復メカニズム

Element fault そこで各 banyan 網の出口に比較器をつけ、banyan 網の出口でパケットの目的地アドレスを出力リンクのラベルと比較して一致しなければ conflict bit がセットされていなくても次の banyan 網に再送するように拡張すれば TBSF で耐故障性を実現できる。当然、コンパレータの故障も考えられる。そのためには、出力ラベルとパケットの 目的地アドレスが一致していても conflict bit がセットされていたら、次の banyan 網に再送すれば良い。この二重チェックを用いることによって、コンパレータの誤動作からも回復することが出来る。

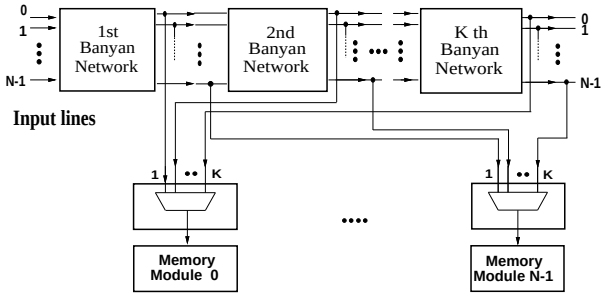

図 2: Tandem Banyan Switching Fabrics

Link fault link fault では、MIN 内でパケットが失われてしまうため、on-the-fly で故障を回復することは難しい。この場合は、スイッチングシステムを停止させ、故障箇所の診断を行ったあと、図 3のように、故障したリンクを含む banyan 網をバイパスする。このため、bypassing path が各 banyan 網に必要である。

故障回復のためにコンパレータとバイパスメカニズムを付加した TBSF を耐故障 TBSF(Fault tolerant TBSF/F-TBSF) と呼ぶ。

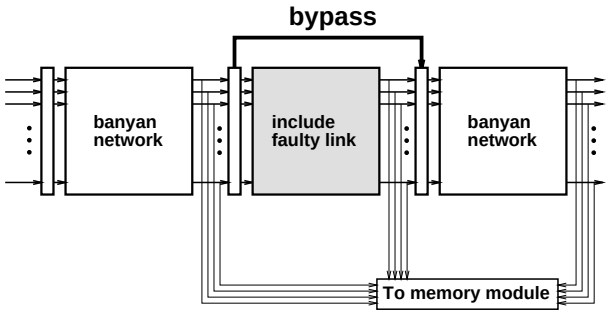

図 3: Bypassing path

3.3 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。この点を改良し、図 4に示すように banyan 網を三次元的に接続した構造に変更した,PBSF(Piled Banyan Switching Fabrics) が提案された。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。

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

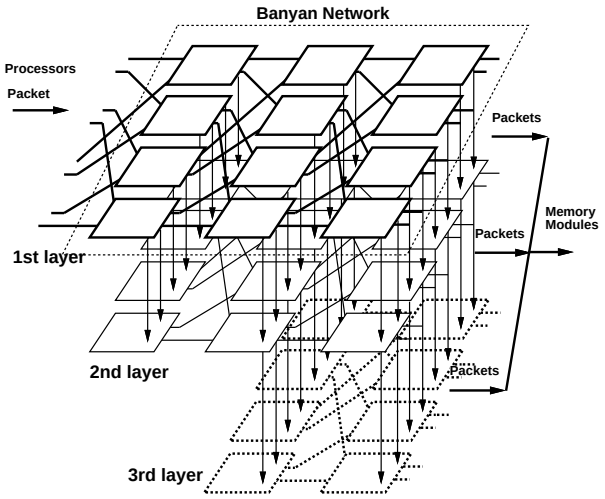

図 4: Piled Banyan Switching Fabrics

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。

3.4 PBSF の故障回復メカニズム

Element fault PBSF でも'elemet fault' の場合、on-the-fly で回復できるが、TBSF と異なり各スイッチングエレメントにチェック機構が必要である。図 5で示すように、垂直方向のリンクは下層の出力リンクと入力リンクに接続する。チェック機構がパケットヘッダのビットをチェックする。もし conflict bit がセットされているならば、普通の PBSF 同様、パケットは下層の出力リンクに転送され、スイッチのステートとヘッダビットが一致しなければ、パケットは下層の入力リンクに転送され再び同じステージのスイッチでルーチングされる。しかし、ここでスイッチのミスルーチングにより下層の入力リンクに転送されたパケットは下層の水平入力と衝突する恐れがある。そこで、各ステージの最下層には水平入力が存在しないことを利用して、下層の入力リンクは全てそのステージでの最下層の入力へ接続する (図 6)。

TBSF に比べてこのメカニズムはハードウェア量を増加させるが、パケットはコントローラとチェック機構が故障しない限りミスルーチングされない。

Link fault PBSF では、故障したリンクはレイヤ間の垂直方向のリンクによりバイパスすることができる。しかしながら、スイッチングエレメントにとって出力リンクが故障しているかどうかを知ることは不可能である。よって TBSF 同様、故障診断と故障箇所の発見は必要である。もし故障したリンクが存在したら、直前のスイッチは故障したリンクをバイパスするために強制的に垂直方向のリンクを使用する。このバイパスモードをスイッチに知らせるために、それ専用のハードウェアか、コントロールパケットが必要である。

上記の故障回復メカニズムを持つ PBSF を耐故障 PBSF(Fault tolerant PBSF/F-PBSF) と呼ぶ。

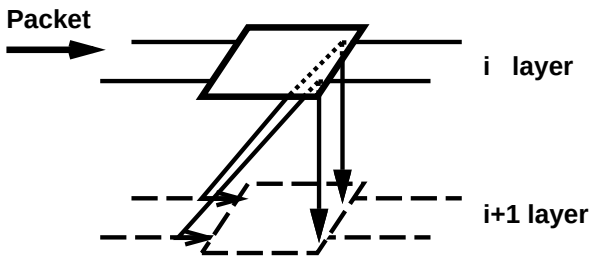

図 5: F-PBSF switching element

4 耐故障性の解析

ここでは、F-TBSF/F-PBSF において故障が存在したときの通過率を確率モデルを用いて解析する。

4.1 F-TBSF における解析

4.1.1 故障が無い場合の TBSF の解析

はじめに banyan 網に故障が存在しないときについて考える。

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時に衝突がおこるのでパケットの衝

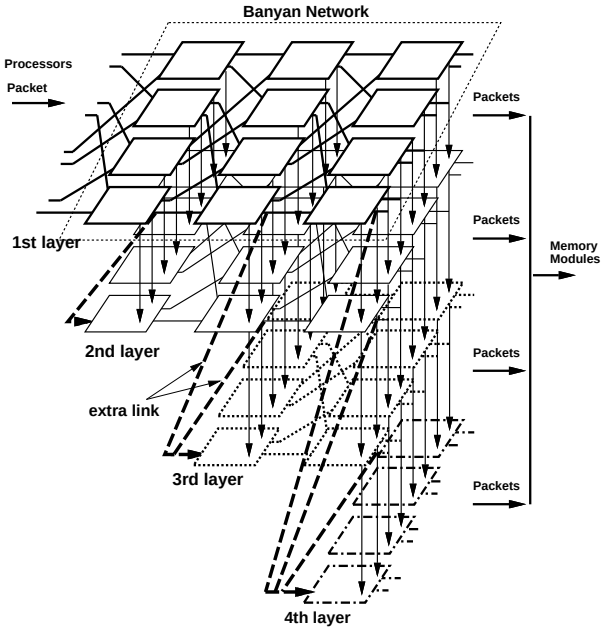

図 6: F-PBSF

突確率は $\frac{1}{4}r^2$ となる。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、即ちパケットは消滅したものと見なされる。

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}は以下のようになる。

$$\begin{aligned} r_{i+1} &= r_i - \frac{{r_i}^2}{4} \\ &= r_i\left(1-\frac{r_i}{4}\right) = r_i f(r_i) \end{aligned} \tag{1}$$

よって banyan 網 1 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を n とすると次のように表される。

$$\begin{aligned} r_n &= r_0\left(1-\frac{r_0}{4}\right)\left(1-\frac{r_1}{4}\right)\cdots\left(1-\frac{r_{n-1}}{4}\right) \\ &= r_0\prod_{j=0}^{n-1}\left(1-\frac{r_j}{4}\right) = r_0\prod_{j=0}^{n-1} f(r_j) \\ &= r_0 f^n(r_0) \end{aligned} \tag{2}$$

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。したがって k番目の banyan 網の各入力にパケットが入力される確率を B_{k-1}とすると次の式が成立する。

$$B_k = B_{k-1} - B_{k-1}f^n(B_{k-1}) \tag{3}$$

で表される。B_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 P_{TBSF}は各 banyan 網における通過率の総和となる [10]。

$$P_{TBSF} = \left(\sum_{k=1}^{l} B_k\right)/B_0 \tag{4}$$

4.1.2 故障が存在する場合の F-TBSF の解析

次に故障が存在する場合について考える。

リンクの故障は banyan 網数 (l) を減らし、通過率の低下は簡単に式 (4) によって解析できる。ここでは、スイッチングエレメントの故障に焦点を当てる。

はじめにスイッチの故障が単一の banyan 網に存在するときを考える。また、故障したスイッチのステージを m ステージとする (図 7)。故障したスイッチでは stuck しているためパケットの衝突は起こらない。そのため見かけの通過率は上がるが、実際は故障したスイッチで間違えて routing されたパケットが存在する。そこでまず見かけの通過率 RA を求め、その後間違えて routing されたパケットの通過率 RMを引くことにより、実際の通過率 RFを求める。

$$RF = RA - RM \tag{5}$$

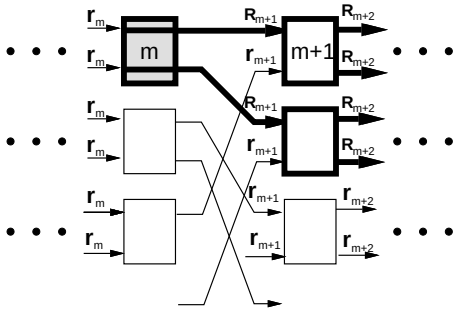

図 7: 故障が存在する場合の各スイッチでの通過率

RA の解析　図 7に示すように、故障したスイッチングエレメントの影響は二分木パスによって広がっていく。

ここで、二分木パス上の入力リンクにパケットが存在する確率を以下のように定める。

r_{i+1}は故障が存在しないときと同様、式 (1)　$r_{i+1} = r_i f(r_i)$ で表される。R_{i+1}は以下の図 8のように考えられる。

よって R_{i+1}は次のようになる。

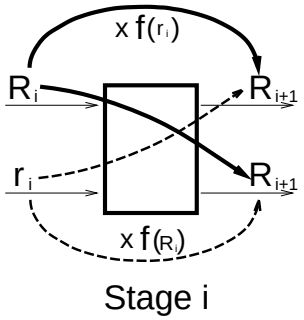

図 8: R_{i+1}の考え方

$$R_{i+1} = \frac{r_i f(R_i) + R_i f(r_i)}{2} \quad (6)$$

故障したスイッチを見かけ上正しく通過したパケットが出力される確率は故障した m ステージでは r_m と等しいが、$m+1$ ステージになると故障したスイッチの影響を受けたパケットが存在するため、r_{m+1} とは異なって来る。よってこれを式 (1), 式 (6) を用いて表すと次のようになる。

$$\begin{aligned} m+1stage &: \frac{2^{n-1}-2^0}{2^{n-1}} r_{m+1} + \frac{2^0}{2^{n-1}} R_{m+1} \\ m+2stage &: \frac{2^{n-1}-2^1}{2^{n-1}} r_{m+2} + \frac{2^1}{2^{n-1}} R_{m+2} \\ &\vdots \end{aligned}$$

よって出力 $(n+1$ ステージ$)$ での見かけの通過した確率 RA は次式になる。

$$\begin{aligned} RA &= \frac{2^{n-1}-2^{n-1-m}}{2^{n-1}} r_n + \frac{2^{n-1-m}}{2^{n-1}} R_n \\ &= \left(1-2^{-m}\right) r_n + 2^{-m} R_n \end{aligned} \quad (7)$$

RM の解析　次に間違えて routing されたまま通過した確率 RM を求める。

m ステージで間違えて routing される確率は

$$\frac{1}{2^{n-1}} r_m \times \frac{1}{2} = \frac{r_m}{2^n}$$

となる。間違えて routing されたパケットが次の $m+1$ ステージを通過する確率はスイッチのそれぞれの出力につき $\frac{r_m}{2^n} f(r_{m+1}) \times \frac{1}{2}$ となる。よって図 7より次のステージに入力する確率は以下のようになる。

$$\frac{r_m}{2^n} f(r_{m+1}) \times \frac{1}{2} \times 2 = \frac{r_m}{2^n} f(r_{m+1})$$

よって間違えて routing されたパケットが出力される確率 RM は次のように表される。

$$\begin{aligned} RM &= \frac{1}{2^n} r_m f(r_{m+1}) f(r_{m+2}) \cdots f(r_{n-2}) f(r_{n-1}) \\ &= \frac{1}{2^n} r_m \prod_{j=m+1}^{n-1} f(r_j) \end{aligned} \quad (8)$$

式 (5), 式 (7), 式 (8) より、m ステージが故障しているときのパケットが出力される確率 RF は以下のようになる。

$$\begin{aligned} RF &= RA - RM \\ &= \left(1-2^{-m}\right) r_n + 2^{-m} R_n \\ &- \frac{1}{2^n} r_m \prod_{j=m+1}^{n-1} f(r_j) \end{aligned} \quad (9)$$

故障したスイッチが $K=1$ の banyan 網に存在する場合の通過率は、$B_1 = RF$ となる。よって故障したスイッチが存在する banyan 網の通過率 F_{FTBSF} は式 (4) より、

$$F_{FTBSF} = \left(RF + \sum_{k=2}^{l} B_k \right) / B_0 \quad (10)$$

となる。

F-PBSF の場合も同様に全ての入力に対する出力の確率を考えて解析する。ページの関係で詳細は省略する。

5　通過率の評価

前章で検討した確率モデルとシミュレーションを用いて F-TBSF/F-PBSF の通過率を解析する。

5.1　F-TBSF

表 1: Pass-through ratio vs. location of the fault on the TBSF (64 inputs, load:0.5, 2 banyan networks)

location of the fault	pass-through ratio	ratio (vs. no fault)
no fault	0.88050	1.00000
stage 0	0.87660	0.99557
stage 1	0.87666	0.99564
stage 2	0.87671	0.99570
stage 3	0.87675	0.99574
stage 4	0.87678	0.99578
stage 5	0.87681	0.99581

図 (9) より、解析結果とシミュレーションの結果は一致しており、解析のためのモデル化が正しいことが分かる。

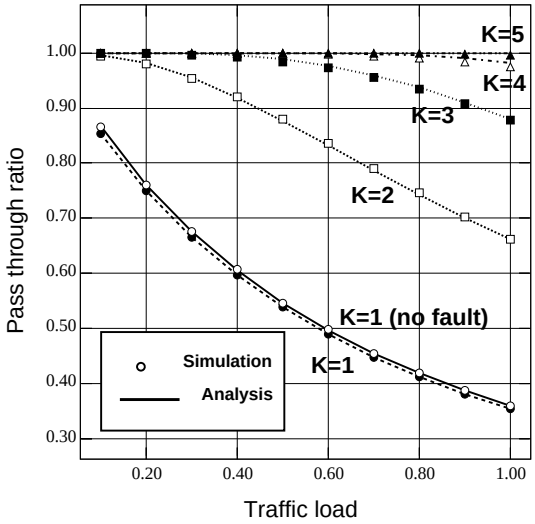

図 9: Pass-through ratio vs. traffic load on the F-TBSF (64 inputs , location of fault: 0 stage)

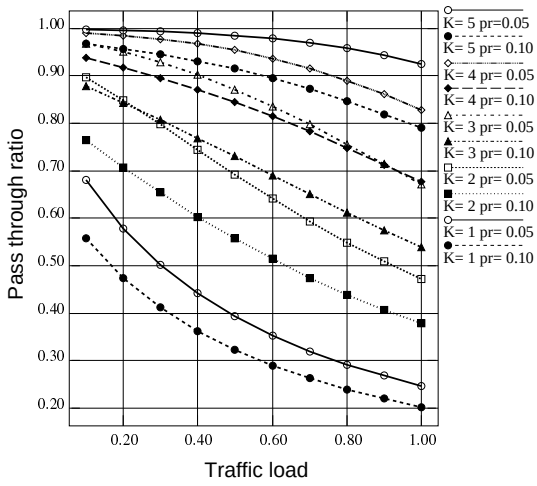

図 10: Pass-through ratio vs. traffic load on the F-TBSF (256 inputs , reliability: 0.95, 0.9)

5.1.1 Element fault

表 1に故障したスイッチが存在するステージと通過率の関係を示した。故障したスイッチが手前のステージに存在する程、通過率は低下する。しかしその影響は小さい (1%以下)。TBSF では、はじめの banyan 網の負荷がもっとも高い。よって、1 番目の banyan 網の 0 ステージが故障しているとき故障の影響は最大となる。

図 (9) はネットワークサイズを 64×64 に固定し、ネットワークに故障したスイッチが存在するときとしないときの入力負荷率 (r_0) に対するパケットの通過率を表している。図 (9) より 0 ステージのスイッチが故障しているときの通過率は故障が無いときに比べ 5 %程度低くなっていることがわかる。また、負荷率を変化させても故障がないときの通過率との比は変化しない。さらに故障したスイッチが 1〜3 ステージに変化しても、表 1より、通過率は若干上がるがほとんど変化しないことがわかった。

図 10はスイッチングエレメントが故障する確率を１０%と５%としたときの入力負荷率に対するパケットの通過率をあらわしている。なお、ネットワークサイズは 256×256 である。故障する確率が 5 %違うだけで、通過率に大きな影響を与えることが分かる。しかしたとえ故障率が１０%もあったとしても、banyan 網を 5 段通過させれば、負荷率が 1.0 のような場合でも８０%以上の高い通過率を保つことが出来る。

5.1.2 Link fault

リンクの故障が存在するとき、banyan 網はバイパスされる。よって、banyan 網の段数 (K) が減ったことになる。図 9に示すように、通過率は Kが小さい程低下する。充分な通過率を補償するには banyan 網は 4 段から 5 段必要である。

5.2 PBSF

F-TBSF と異なり F-PBSF では故障回復メカニズムは各スイッチに付加されている。よって、パケットはスイッチングエレメントの故障、リンクの故障の両方の場合にパケットは下層に送られる。よって、両方の故障の場合の通過率の低下は同じである。PBSF では TBSF と同様、1 段目の banyan 網のステージ 0 のスイッチが故障したときがもっとも故障の影響が大きい。

図 (11) にネットワークサイズを 256×256、スイッチが故障する確率を１０%に固定し、PBSF の layer 数を変化させたときの入力負荷率に対するパケットの通過率を表す。layer 数が 3 以上になると解析結果とシミュレーションの結果に数%の誤差がある。これは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が、PBSF では前の layer によってトラフィックの均一さが失われるためである。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。

図 (11) より故障率が１０%だと、1 段目の通過率は 5〜10 %減少する。また、layer 数 Kを増加させると負荷率が 1.0 のときは通過率が 28 %程度だが 2 段通過した後は 55 %以上、3 段通過した後は 75 %以上になることが分かった。これは TBSF よりも高い通過率を示している。

図 10は F-TBSF 同様スイッチングエレメントが故障する確率を１０%と５%としたときの入力負荷率に対するパケットの通過率をあらわしている。この図より、banyan 網を 3 段通過させれば、どちらの故障率でも６５%以上の通過率を保つことが出来る。また、banyan 網を 3 段通過した後は、故障率の影響が少なくなることがわかる。

図 (10) と図 (12) より、接続されている banyan 網の段数が低いとき (2,3 段)F-PBSF の通過率は F-TBSF の通過率より 10 %高いことがわかる。PBSF では、衝突するま

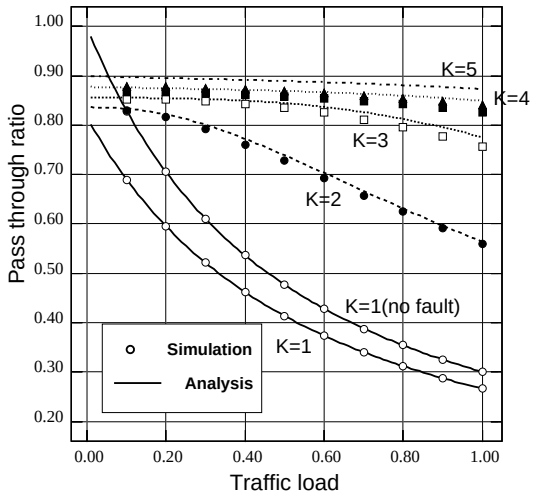

図 11: Pass-through ratio vs. traffic load on the PBSF(256 inputs , reliability: 0.95)

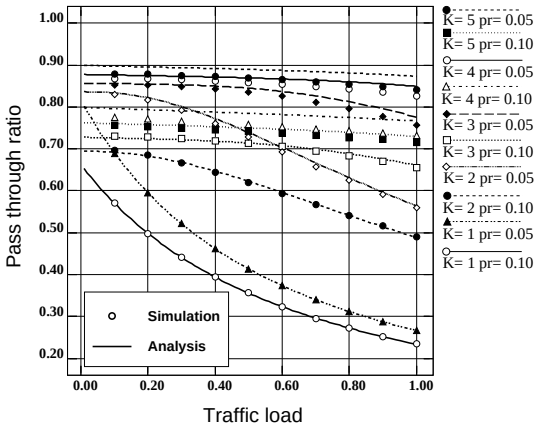

図 12: Pass-through ratio vs. traffic load on the PBSF(256 inputs , reliability: 0.95, 0.90)

でのルーチングは無駄にはならない特徴を持ち、TBSF に比べて有利なためである。

6 結論

TBSF、PBSF という多重出力可能な MIN の故障回復メカニズムを提案した。これらのメカニズムを用いることにより、スイッチングエレメントの複数の故障に対して on-the-fly で故障回復を行うことが出来る。リンクの故障のときは、故障診断の後、ネットワークは再構成され再び使用できる。

スイッチが複数故障した場合の通過率を確率モデル、シミュレーションによって解析した。PBSF は故障率が高く、負荷率が高いような場合でも、3 段の banyan 網を通過することで優れた耐故障性を示した。

参考文献

[1] H.Amano, L.Zhou, K.Gaye, "SSS(Simple Serial Synchronized)-MIN: a novel multi stage interconnection architecture for multiprocessors," Proc. of the IFIP 12th World Computer Congress, Vol.I, pp.571-577, Sept. 1992.

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

[3] F.A.Tobagi, "Fast Packet Switch Architectures For Broadband Integrated Services Digital Networks," Proceedings of the IEEE Vol.78, No.1 Jan. 1990.

[4] H.Sakamoto, T.Masaki, H.Inoue, H.Amano, "Configuration and evaluation of self routing switches," ISSE88-30 No.8, 1988, (in Japanese).

[5] F.A.Tobagi and T.Kwok, "The Tandem Banyan Switching Fabric: a Simple High-Performance Fast Packet Switch," Proc. INFOCOM91, Apr. 1991.

[6] C.L.Wu, M.Lee, "Performance Analysis of Multistage Interconnection Network Configurations and Operations," IEEE. Trans. Comput., Vol. 41, No.1 pp.18-27, Jan. 1992.

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

[8]

[9] Tse-Yun Feng, "Fault Diagnosis for a Class of Multistage Interconnection Networks", IEEE Trans. on Computer C-30, 10, pp.351-366 (Oct. 1981).

[10] T. Hanawa, H.Amano, Y.Fujikawa, "Multistage Interconnection Networks with multiple outlets," Proc. of International Conference on Parallel Processing, Vol.I pp.1-8 (Aug. 1994).

[11] M.Sasahara, J.Terada, L.Zhou, K.Gaye, J.Yamato, S.Ogura, H.Amano, "SNAIL: a multiprocessor based on the Simple Serial Synchronized multistage interconnection network architecture," Proc. of International Conference on Parallel Processing, Vol.I pp.76-80 (Aug. 1994).

[12] N.J.Davis IV, W.T.Hsu, H.J.Siegel, "Fault location techiniques for Distributed Control Interconnection Networks," IEEE Trans. on Computer C-34, 10, pp.902-910 (Oct. 1985).

[13] A. Jajszczyk J,Tyszer, "Fault Diagnosis of Digital Switching Networks," IEEE Trans. on Communication, COM-34, 7, pp.732-739, (July 1989).