自律再構成可能な格子結合型マルチプロセッサシステムのハードウェア構成

山田 順也　　阿部 亨　　堀口 進
北陸先端科学技術大学院大学 情報科学研究科
〒923-1292 石川県能美郡辰口町旭台 1-1
E-mail: {j-yamada,beto,hori}@jaist.ac.jp

あらまし　近年，大規模な計算を高速に処理するために，多数の Processing Element（PE）を相互結合網で接続した超並列計算機に関する研究が盛んに行われている．超並列計算機には多数のプロセッサが存在するため，プロセッサの故障回避を考慮したシステムのフォールトトレランスが重要な問題である．超並列計算機の一つである格子結合型マルチプロセッサのプロセッサ故障回避として，トラックとスイッチを用いた再構成手法がいくつか提案されている．しかし，従来の再構成手法はアルゴリズムの解析に重点がおかれたシミュレーション評価に留まっており，ハードウェアへの実装については考慮されていない．本稿では，自律再構成が可能な格子結合型マルチプロセッサシステムのハードウェア実装について検討する．また，自律再構成が可能な格子結合型マルチプロセッサシステムの再構成時間，冗長な回路量について検討し，FPGA 上に実装した場合の評価を行う．

キーワード　格子結合型マルチプロセッサ，自律再構成，故障回避，フォールト トレランス，FPGA

A Hardware System of Self-Reconfigurable 2D-Mesh Multiprocessor

Junya Yamada　　Toru Abe　　Susumu Horiguchi

School of Information Science
Japan Advanced Institute of Science and Technology
1-1 Asahi-Dai, Tatsunokuchi, Nomi, Ishikawa, 923-1292, Japan
E-mail: {j-yamada,beto,hori}@jaist.ac.jp

Abstract　To achieve high-performance computing for large scale advanced applications, many researchers have been studying massively parallel computers consisting of a large number of processing elements(PEs). A Fault-tolerance is one of the critical problems to construct massively parallel computers. Several reconfiguration architectures using spare PEs, tracks and switches have been proposed for 2D-mesh multiprocessor systems. Several reconfiguration schemes of 2D-mesh array theoretically discussed by only simulations. However, a hardware implementation of the reconfiguratoion algorithms to achieve fault-tolerance have not been studied yet. We propose new hardware systems to achieve self-reconfiguration of 2D-mesh multiprocessor systems. The hardware systems are implemented on FPGAs and evaluated on reconfiguratoion times and the number of gates required for redundant circuits.

key words　2D-mesh Array, Self-Reconfiguration, Defect Avoidance, Fault Tolerance, FPGA

1 はじめに

近年，VLSI 技術の目覚ましい進歩に伴いコンピュータの性能は飛躍的に向上した．しかし，現在のスーパーコンピュータをもってしても最先端の科学技術分野における大規模シミュレーションには膨大な計算時間が必要となる．そこで，高速計算を行うため，超並列システム，シストリックアレイ，ニューロコンピュータ等の研究が盛んに行われている．超並列システムでは，システムを構成する PE(Processing Element) の数が膨大であり，故障 PE の回避を含めたフォールトトレランスが必要不可欠である．

本稿では，超並列システムの一つである格子結合型マルチプロセッサシステムに着目し，自律再構成が可能な格子結合型マルチプロセッサシステムのハードウェア実装について検討する．また，自律再構成が可能な格子結合型マルチプロセッサシステムを FPGA(Field Programmable Gate Array) 上に実装し，再構成時間，冗長な回路量を含めた総合的な評価を行う．

2 格子結合型マルチプロセッサシステム

2.1 冗長構成

格子結合型マルチプロセッサシステムに対するフォールトトレランスは，現在までに様々なものが提案されている．M.Chean ら [1] によると，主に時間冗長とハードウェア冗長の 2 つに分類される．

時間冗長とは故障が発生した場合，ある PE が故障 PE の代用を行うものである．R.Negrini ら [2] は故障 PE がある場合にスイッチ回路を切り替え，再度処理を PE で行う時間冗長アーキテクチャを提案している．しかし，時間冗長では，クロックの遅延が生じて PE 間の同期が崩れるため，演算速度に対する影響が大きい．そのためパイプライン的な処理を行う高速動作には一般に不向きである．しかし，付加する回路が小規模で済む利点がある．

ハードウェア冗長はシステムに冗長な PE を付加し，アレイ上に故障が発生した場合スイッチを切り替えて，故障 PE を回避するアーキテクチャである．ハードウェア冗長では付加する冗長な回路が大きいが，PE 間接続の遅延を抑えることができる．しかし，システムに付加する冗長な PE により故障箇所が増加するという問題点がある．

2.2 再構成システム

S.Y.Kung ら [3][4] は，$N \times N$のアレイに 1 行 1 列または 2 行 2 列の冗長 PE を付加し，スイッチ回路を用いることにより $N \times N$のアレイを得るアーキテクチャを提案している．これは各 PE 間にそれぞれ 1 本または 2 本のトラックおよびスイッチ回路があり，故障の分布によって各スイッチの切り替えを行う再構成システムである．

このシステムでは，再構成問題は $(N + 2R) \times (N + 2R)$ のアレイから結合制限を満たす $N \times N$ のアレイを得ることと定義できる．ここで，n 個の PE が故障しているとするとスイッチ切り替えの組み合わせは 4^n となり，n が増加するにつれて指数的に増加する．これらの組み合せを一つ一つ調べて探索することは，現在のコンピュータでは非常に困難である．

S.Y.Kung ら [3] は，グラフ理論を用いて探索の数を減らしている．しかし，冗長な PE の数が増えるとスイッチ切り替えの組み合せはさらに増加する．2 行 2 列の冗長な PE を付加した場合は，全ての組み合せを調べるのは困難である．そこで，S.Y.Kung ら [4] は，PE 間のトラックの数を 2 本として、PE 間に仮の PE を一段置くことにより 1 行 1 列の場合に近似させてこの問題を解いている．これら 2 つのアルゴリズムは無駄なく PE を有効に利用することができるが、1 行 1 列にのみにしか有効でない．したがって，システム上の全故障情報を用いて故障 PE を回避するようにスイッチを切り替える方法は，システムの規模が大きくなるにつれ故障個所を検出したり，外部からスイッチの切り替えを行うのが困難となる．

これらに対して沼田ら [5][6] は，各 PE が自分を含めた近隣 PE の故障情報のみで自律再構成を行う手法 (BS, FS, HS) を提案している．これらの手法は，グラフ理論を用いた再構成手法と同程度の再構成率を得ることができる．しかし，バイパス方向を乱数で決定したり隣接 PE 間のメッセージ通信が必要であるなど実装上の問題点については十分に検討されていなかった．

そこで本稿では，二次元格子結合型マルチプロセッサを対象とした，ハードウェア実装が容易な自律再構成アルゴリズムを新たに提案する．

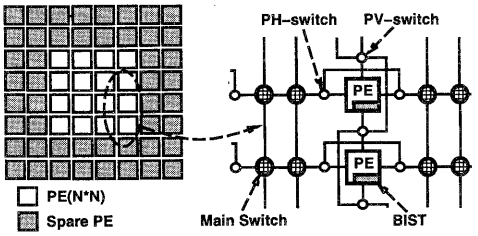

図 1 格子結合型マルチプロセッサシステム構成

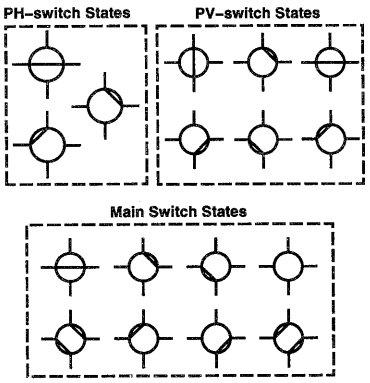

図 2 各スイッチ状態

3 自律再構成法

3.1 システム構成

図1に格子結合型マルチプロセッサシステムの概略を示す．提案するシステムは，$N \times N$の格子結合型マルチプロセッサ，その周辺に配置されたR行R列の冗長なPE，PE間のR本のトラックとスイッチ回路から構成される．周辺部にある冗長なPEと中心部にあるPEは同等で，機能，性能ともに違いはない．故障を含む $(N+2R) \times (N+2R)$ のアレイから故障PEをスイッチ回路で回避して $N \times N$ のアレイを再構成する．各PEは故障を検知する自己テスト回路 (Built In Self Test Circuit :BIST) を持ち，各スイッチ回路は周辺PEのローカル故障情報のみでスイッチの切り替えを行うシステムである．自律再構成システムでは，左右のPE間接続を行うメインスイッチ，水平方向のバイパスに使用される PH-スイッチ，垂直方向のバイパスに使用される PV-スイッチが各PEに付加される．

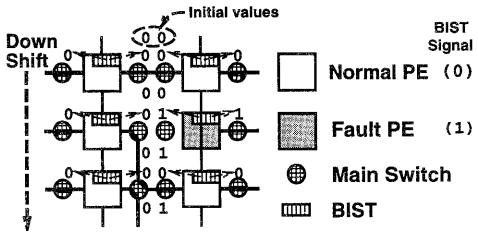

図 3 Main Switch の接続決定方法

3.2 スイッチ回路

図2に各スイッチがとりうる状態を示す．メインスイッチは，左右のPEから故障信号を受けて接続状態の選択を行う．PH-スイッチは，上下のPEから故障信号を受けて接続状態の選択を行う．PV-スイッチは，付加されているPEの故障信号を受けて接続状態の選択を行う．

図3にメインスイッチの詳しい接続決定方法を示す．各PE内のBISTは，PEが故障していれば (1)，正常であれば (0) の信号を左右のメインスイッチへ送る．各メインスイッチはその信号と上位のメインスイッチから送られてきた信号によって接続状態が決定される．接続状態が決定すると，BISTからの故障信号と上位のメインスイッチからの信号を加算し，その結果を下位のメインスイッチへ渡す．この処理を各列ごとに繰り返して行い，故障PEを回避した結合網を再構成する．

3.3 自律再構成アルゴリズム

3.3.1 2wBS(2way Bypass and Shift) 法

バイパスと2方向のシフトを組み合わせた2wBS自律再構成アルゴリズムを図4に示す．

このアルゴリズムは，行方向の回避にバイパスを用いるので，行方向のスイッチ回路およびトラックを必要としない．そのため，付加する冗長な回路が小さくなる．

3.3.2 4wBS(4way Bypass and Shift) 法

2wBS法より高い歩留まりを得るために，2wBS法に行方向のシフトを加えた4wBS自律再構成アルゴリズムを図5に示す．

4wBS法は，列方向と行方向にスイッチ回路とトラックを必要とするため，付加する冗長な回路と配線が2wBS法より多くなる．

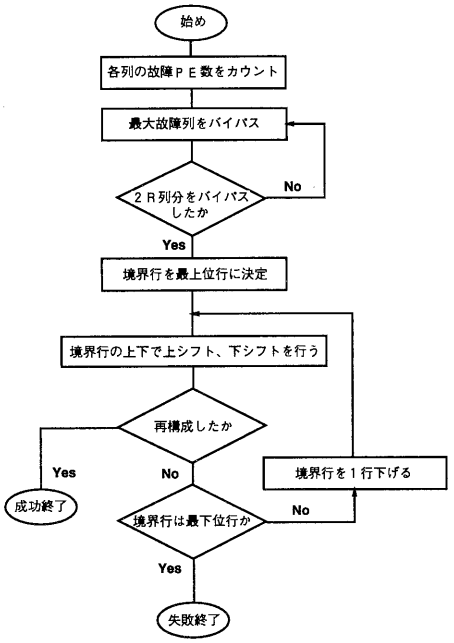

図 4 2wBS 法

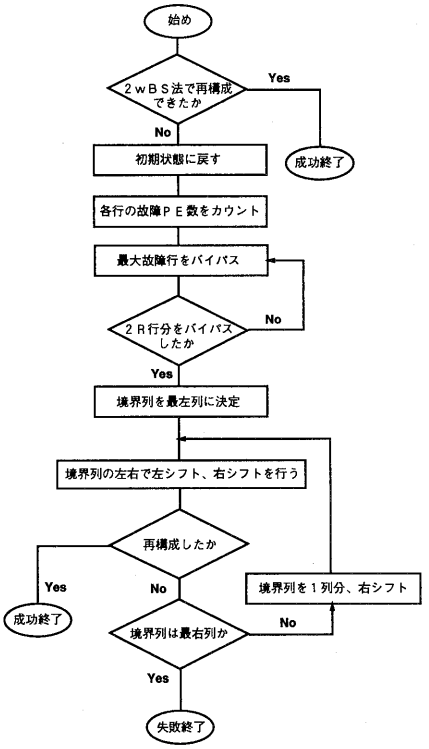

図 5 4wBS 法

4 自律再構成法の性能評価

4.1 アレイ歩留まり

格子結合型マルチプロセッサシステムのアレイ歩留まり (Array Yield) について性能評価を行う．本稿におけるアレイ歩留まりはアレイが再構成できる確率として定義し，PE 歩留りは全システム内における正常な PE の生存確率として定義する．従来の再構成システム [3]～[6] と比較するために下記の仮定を行う．

- 故障を検知する回路は故障しない．
- トラックおよびスイッチ回路は PE の規模に比べ極めて小さいので故障しない．
- 各 PE は一定確率で故障する．
- 故障箇所はランダムかつ一様に分布する．

各故障数につき 10 万回のシミュレーションを行い，平均のアレイ歩留まりを求めた．

4.1.1 2wBS 法の歩留まり評価

図 6に，2wBS 法による 8×8 と 16×16 の基本アレイ周辺に冗長な PE を加えた場合の平均アレイ歩留まりを示す．図中の 2wBS(8+2) は，8×8 の基本アレイに 1 行 1 列の冗長な PE を付加した場合で，2wBS(8+4)，2wBS(8+6) は，2 行 2 列，3 行 3 列の冗長な PE を付加したことを示す．アレイサイズが小さい時は，1 行 1 列の冗長な PE でも比較的高いアレイ歩留まりが得られている．しかし，16×16 の基本アレイに対しては冗長な PE が 1 行 1 列ではアレイ歩留まりが極端に悪くなる．

図 7に，冗長な PE を 2 行 2 列付加し，基本アレイサイズを変化させた場合のアレイ歩留まりを示す．図 7より，基本アレイサイズが大きくなるとアレイ歩留まりが悪くなる．

図 8に，2wBS 法と従来法とのアレイ歩留まりの比較を示す．2wBS 法は，基本アレイサイズが 10×10 程度であれば Kung の手法よりも優れているが，基本アレイサイズが 20×20 程度になると Kung の手法より歩留まりが低くなる．その理由として，2wBS 法では大規模なアレイサイズの最大故障列をバイパスする時に，多数の正常な PE をも切

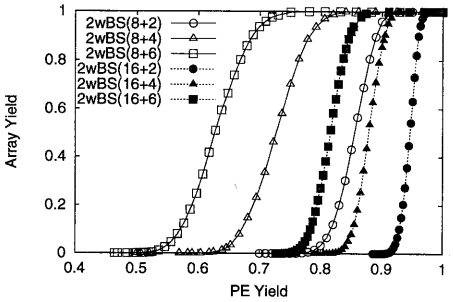

図 6 冗長な PE 数を変化させた場合のアレイ歩留まり

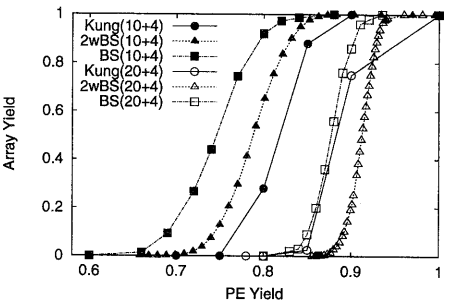

図 8 従来法とのアレイ歩留まりの比較（2wBS）

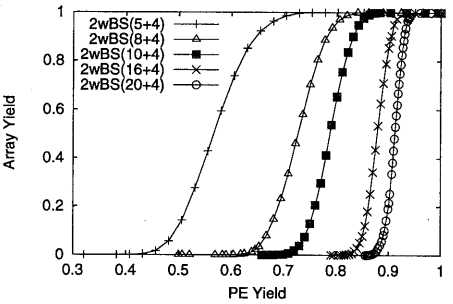

図 7 アレイサイズを変化させた場合の各歩留まり

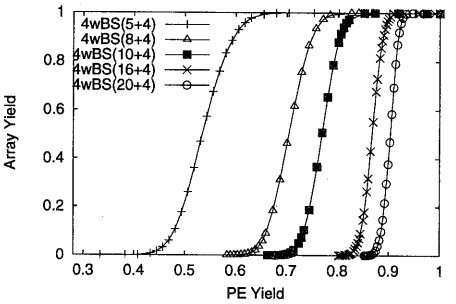

図 9 アレイサイズを変化させた場合の各歩留まり（4wBS）

り離してしまうことと，ハードウェアによる単純な上下シフトしか行わないことが挙げられる．

4.1.2 4wBS 法の歩留まり評価

2wBS 法では，アレイ歩留まりの性能が従来法と同程度になることが分かった．次に 2wBS 法に行方向へのシフトを加えた 4wBS 法のアレイ歩留まりについて評価を行う．図 9に，冗長な PE が 2 行 2 列で基本アレイサイズを変化させた場合のアレイ歩留まりを示す．また図 10 に，4wBS 法と従来法のアレイ歩留まりの比較を示す．4wBS 法は 2wBS 法より高いアレイ歩留まりを得ることが分かった．次に，冗長な PE を 2 行 2 列にした場合で，2wBS 法と 4wBS 法のアレイ歩留まり性能について詳しく検討する．

図 11は，2wBS 法と 4wBS 法のアレイ歩留まりを比較したものである．図 12に，R=2 とした場合の 2wBS 法と 4wBS 法の各基本アレイサイズにおけるアレイ歩留まり性能を示す．アレイ歩留まりの性能は各アレイサイズにおける全故障発生パターンのなかから再構成が可能であった平均確率である．

故障 PE 数が少ないうちは 4wBS 法の方が常にアレイ歩留まりがよい．これは 2wBS 法で再構成できない場合も 4wBS 法では行方向へのシフトによって再構成できるからである．しかし，再構成不可能になる故障 PE 数についてはほとんど同数である．また，基本アレイサイズが 4×4 の場合，2wBS 法から 4wBS 法に改善することで得る歩留まりの向上率は 8%と高いが 20×20 になると向上率は 1%となり，大規模なアレイサイズになるほど歩留まりの向上率が低下している．これは，4wBS 法は単純に 2wBS 法を列方向と行方向に繰り返したアルゴリズムのためであり，大規模なアレイサイズになると 2wBS 法と同程度の歩留まりになる．

4.2 再構成時間

次にシステムの再構成時間について検討する．ここで，各処理時間を以下のように定義する．

T_{SW}: 1PE につながる全スイッチの切替え時間．

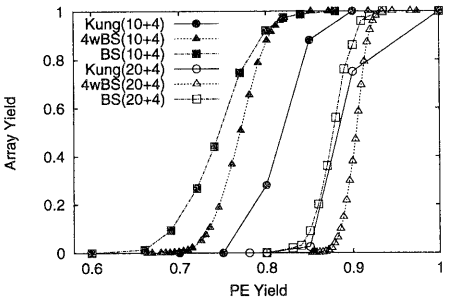

図 10 従来法とのアレイ歩留まりの比較（4wBS）

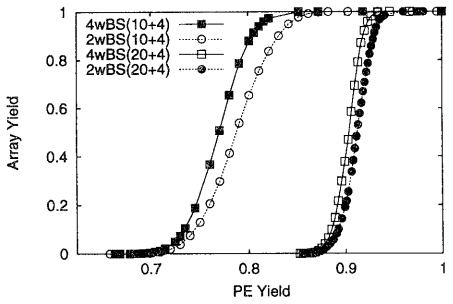

図 11 4wBS 法と 2wBS 法のアレイ歩留まり比較

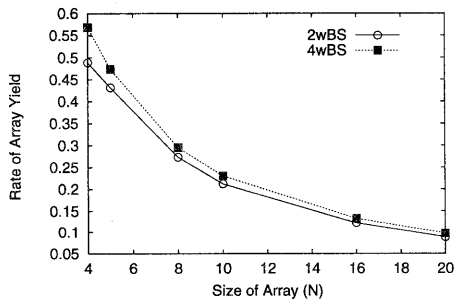

図 12 アレイ歩留まりの性能比較，R=2

表 1 最小再構成時間の比較，R=2

タイプ	最小再構成時間
Kung	$2(N+4)T_{SW}+2T_{I/O}+T_{Host}$
BS	$(N+3)(N+4)T_{SW}$
2wBS	$4(N+4)T_{SW}$
4wBS	$4(N+4)T_{SW}$

表 2 最大再構成時間の比較，R=2

タイプ	最大再構成時間
Kung	$2(N+4)T_{SW}+2T_{I/O}+T_{Host}$
BS	$3(N+4)T_{SW}+8^{N(N+4)}T_{SW}$
2wBS	$3(N+4)T_{SW}+(N+4)^2T_{SW}$
2wBS	$6(N+4)T_{SW}+2(N+4)^2T_{SW}$

$T_{SW} << T_{I/O} << T_{Host}$

$T_{I/O}$: チップとホストマシン間のデータ転送時間．

T_{Host}: ホストマシン内での計算時間．

表 1，表 2に $N \times N$の基本アレイに 2 行 2 列の冗長な PE を付加したシステムを各手法で再構成を行った場合の最小，最大再構成時間をそれぞれ示す．BS 法 [6]，2wBS 法，4wBS 法はローカルな故障 PE 情報のみで自律再構成を行うので，システム全体のグローバルな故障 PE 情報を用いた Kung[4] の手法に比べて再構成時間は短い．

また，BS 法はスイッチ状態の決定に，非決定アルゴリズムを用いているので 2wBS 法，4wBS 法より再構成時間がかなり長くなる．これは，BS 法がシステム全体の PE を何周も巡回しながら徐々に再構成を行うのに対して，2wBS 法と 4wBS 法は各列または各行ごとに並行して再構成を行うからである．4wBS 法は 2wBS 法を基本としているので最小再構成時間については同じになるが，最大再構成時間では行方向のシフトを行う時間が余分に必要になる．したがって，2wBS 法が最も高速な再構成手法である．

4.3 総合評価

表 3に，格子結合型マルチプロセッサの各再構成法の歩留まり，ハードウェア量および再構成時間について評価した結果を示す．

BS 法は，全体的に優れているが実装に不向きという問題点がある．4wBS 法は 2wBS 法に比べて冗長な回路量が増えるにもかかわらず，歩留まりはあまり向上していない．したがって，実装の容易さからは 2wBS 法が最も優れていると言える．

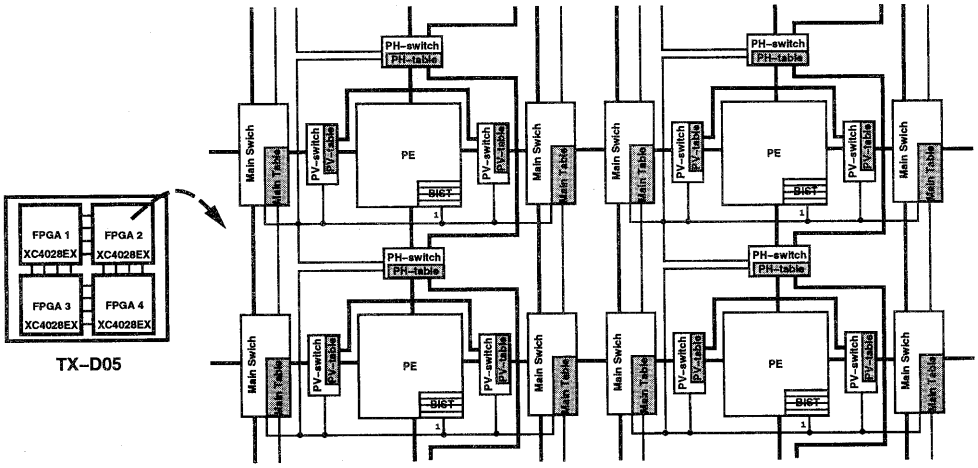

図 13 FPGA 内のブロック図

表 3 再構成手法の比較

タイプ	歩留まり	ハードウェア量	再構成時間
Kung	○	×	×
BS	◎	◎	△
2wBS	○	◎	◎
4wBS	○	×	○

5 ハードウェア実装

5.1 実装環境

本稿で提案した格子結合型マルチプロセッサの自律再構成法をハードウェア上に実装する．実装環境として，以下のソフトや FPGA を使用する．なお，回路設計には VHDL(Very Hard Description Language) を用いる．

- IBM PC/AT 互換機
- 回路設計ソフト (Foundation 1.4 : XILINX)
- FPGA (XC4028EX : XILINX)
- FPGA ボード (TX-D05 : Towa Elex)

5.2 ハードウェア構成

図 13に，2 次元格子結合網に 2wBS 法を実装したときのブロック図を示す．FPGA ボード上の各 FPGA には，PE，メインスイッチ，PH-スイッチ，PV-スイッチが格子結合網に接続されている．2wBS 法では各列ごとに故障数をカウントするので，最下位行が存在する FPGA3 と FPGA4 には加算器や比較器が余分に必要となる．

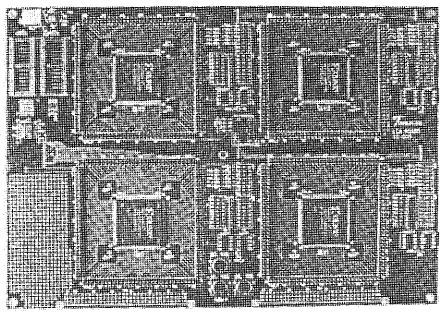

図 14 FPGA ボード (TX-D05)

PH-スイッチ，PV-スイッチは，BIST からの故障信号で各テーブルを参照し，決められたスイッチの切り替えを行う．メインスイッチは，BIST からの故障信号に上位のメインスイッチからの制御信号を加えた信号でメインテーブルを参照し，決められたスイッチの切り替えを行う．なお，本稿では BIST については議論しないので BIST からの信号には GND，VCC を用いる．

図 14に FPGA ボードを示す．本ボードは 4 つの FPGA を接続しており合計 10 万ゲート相当の大規

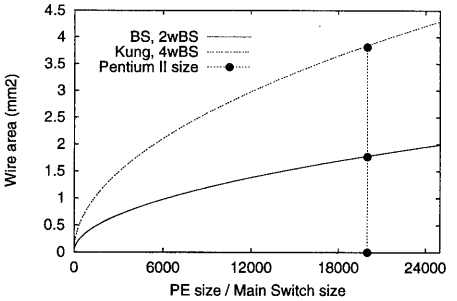

図 15 PE サイズと配線面積の関係

模な回路を実装できる．

図 15に Pentium II プロセッサにメインスイッチを付加した場合の配線量を示す．この図よりメインスイッチの規模はプロセッサの約 2 万分の 1 と小さく，プロセッサに比べて故障する確率は非常に小さい．また，BS 法や 2wBS 法の配線量は，Kung の手法や 4wBS 法のほぼ半分ですむことが分かる．これは，BS 法や 2wBS 法では行方向のシフトがないために，行方向の配線やスイッチ回路がなく，チップ面積が小さくなる．

6 おわりに

本稿では超並列システムとして実現が期待されている格子型結合マルチプロセッサシステムの自律再構成法と実装について議論した．自律再構成型マルチプロセッサシステムは基本プロセッサアレイ，冗長プロセッサ，自己テスト回路，スイッチおよびトラックから構成される．

従来のグローバルな故障 PE 情報を用いて再構成を行う方法や非決定アルゴリズムを用いた自律再構成法に対して，本稿ではハードウェア実装を考慮して，ローカルな故障 PE 情報のみで自律再構成が可能な手法を提案した．これらの自律再構成法により，グラフ理論を用いた再構成法と同程度の歩留まりが得られ，高速な再構成を実現できることを明らかにした．

現在，メッシュ結合システムをトーラス結合システムに拡張した自律再構成について検討している．

謝辞

なお，本研究は文部省科学研究費 基礎研究 (B) ならびに国際学術研究 (共同研究) を用いて行われた．関係各位に感謝する．

参考文献

[1] M.Chean, Jose A.B.Fortes: "A Taxonomy of Reconfiguration Techniques for Fault-Tolerant Processor Arryas", *IEEE Computer*, Vol.23, No.1, pp.55-69 (Jan. 1990)

[2] R.Negrini, R.Stefanelli: "Time Redundancy in WSI Arrays of Processing Elements", *Proc. 1st Int. Conf. on Supercomputing Systems, St.Petersburg* (Dec. 1985)

[3] S.Y.Kung, C.W.Chan: "Fault-tolerant array processors using single-track switches", *IEEE Trans. Computers*, Vol.38, No.4,pp501-514 (Apr. 1989)

[4] J.S.N.Jean, H.C.Fu and S.Y.Kung: "Yield enhancement for wsi array processors using two-and-half-track switches", *IEEE Int'l Conf.on Wafer Scale Integration*, pp.243-250 (Jan. 1990)

[5] 沼田一成, 堀口 進: "格子結合型マルチプロセッサシステムの自律再構成法", 電子情報通信学学会論文誌 (D-I),J76-D-I,pp.531-540 (Oct. 1993)

[6] 沼田一成, 堀口 進: "格子結合型マルチプロセッサシステムの WSI 構成法", 電子情報通信学学会論文誌 (D-I),J77-D-I,pp.121-129 (Feb. 1994)

[7] S.Horiguchi: "Systolic sorter for WSI implementation", *Proc. IEEE Int'l Conf. Wafer Scale Integration*, pp.151-160 (Jan. 1989)

[8] S.Horiguchi, I.Numata, M.Kimura: "Self-Reconfigurable Algolithm of WSI Sorting Network", *IEEE Int'l Conf. on Wafer Scale Integration*, pp.249-255 (Jan. 1991)

[9] 高浪 五男, 久長 穣, 井上 克司: "周辺に予備を持つ格子状結合並列計算機の再構成のニューラル・アルゴリズム", 信学技報 (WSI92-7), pp126-131 (1992)

[10] 高浪 五男, 松野 浩嗣: "WSI におけるフォールトトレランス", 電子情報通信学学会誌, pp888-892 (Sep. 1998)