ハイパフォーマンス
コンピューティング 68－1
（1997．10．17）

SRT 相互結合網のウェーハスタック実装における冷却について

井口 寧†, 松澤 照男†, 堀口 進‡

† 北陸先端科学技術大学院大学 情報科学センター
‡ 北陸先端科学技術大学院大学 情報科学研究科

〒 923-12 石川県能美郡辰口町旭台 1 丁目 1 番地
TEL: 0761-51-1306, 0761-51-1301, 0761-51-1265
e-mail: inoguchi@jaist.ac.jp, matuzawa@jaist.ac.jp, hori@jaist.ac.jp

本論文では，超並列計算機用の再帰シフトトーラス (SRT) 相互結合網の WSI スタック実装時における冷却について議論する．WSI スタック実装は，大規模な超並列システムの実装方式の一つであるが，ウェーハ上の欠陥の回避や中心近くのプロセッサの冷却などが大きな問題となる．SRT 網は超並列システムに適する階層的な相互結合網であり，WSI 実装時に必要とされる欠陥回避が可能な結合網である．そこで，放熱を効果的に行ないつつ欠陥を回避する，WSI スタック実装 SRT の欠陥回避方式を提案する．本方式について，システム全体の歩留まりと WSI スタック内の最高温度をシミュレーションにより求めたところ，冗長プロセッサをウェーハ周囲に配置することにより WSI スタックを効果的に冷却できることが分った．

キーワード：相互結合網，超並列システム，WSI スタック，フォールトトレラント，欠陥回避．

Cooling Methods for SRT Interconnection Network on 3D Stacked Implementation

Yasushi Inoguchi†, Teruo Matsuzawa†, Susumu Horiguchi‡

† Japan Advanced Institute of Science and Technology, Center for Information Science
‡ Japan Advanced Institute of Science and Technology, School of Information Science

Tatsunokuchi, Ishikawa 923-12. Japan.
TEL: +81-761-51-1306, +81-761-51-1301, +81-761-51-1265
e-mail: inoguchi@jaist.ac.jp, matuzawa@jaist.ac.jp, hori@jaist.ac.jp

This paper addresses the reconfiguration of Shifted Recursive Torus (SRT) network by considering thermo-radiation in stacked wafers implementation. The SRT networks are hierarchical torus networks and suitable for massively parallel systems. We propose fault-tolerance schemes for SRT networks to keep highly network performance in stacked wafers implementation. The cooling of stacked wafers, however, is one of the most crucial problems for implementation massively parallel systems. Two cooling approaches have been proposed for SRT in stacked implementation. Introducing a thermo-radiation model into SRT in stacked implementation, reconfiguration performance of SRT was evaluated. Comparing the system yields and the maximum temperatures, these cooling approaches can keep high system yield and lower temperature of 3D implementation.

Keywords: Interconnection Network, Massively Parallel Computer, WSI, Wafer Stack Integration, Fault Tolerant, Defect Avoidance.

1 はじめに

多数のマイクロプロセッサ (MPU) を結合した並列計算機は，自然科学分野におけるシミュレーションなどの多くの分野で，大規模科学技術計算を高速に実行することができる新しい計算機として大きく期待されている．並列処理を行なう場合，ノードとなるプロセッサ要素 (PE) 間通信が処理能力全体を規定する重要な要素となるため，様々な特徴を持つ相互結合網が提案されている [1]．格子結合やトーラス結合は，トポロジが科学技術計算に良く適合し，実装が容易であるという特徴を持つが，規模の拡大に従って網の直径が大きくなり，大規模なシステムには適さない．そこで，格子結合に複数のバイパスリンクを設けることによりネットワーク性能を向上させた様々な相互結合網が提案されている [2, 3]．先に筆者らが提案した 2 次元 SRT 網 [4] は，トーラス網を基本としてグリッド間隔を 2 の整数乗倍しながら階層的に重ね合せて構成される相互結合網であり，小さい直径や平均距離を少ないリンクで実現している．

一方，実装の観点から，1 枚のウェーハ上に多数のプロセッシング要素 (PE) を搭載するウェーハスケールインテグレーション (WSI) が注目されている．超並列システムを WSI により構築した場合，PE 間の相互結合網が全てチップ内部で可能となるため，クロックの遅延が少なくなり，システムの小型化，高速化，小電力化が期待できる．しかしながら，大口径ウェーハ上に発生する欠陥は，現在の集積技術では避けられない問題である．このため，WSI 技術により超並列システムを構築するためには，欠陥の回避が可能な相互結合網を採用する必要がある．WSI システムの欠陥回避は，格子結合網について多くの研究がなされている [5, 6]．これらの多くでは，初期に動作すると仮定している PE の周囲に冗長 PE を配置し，故障 PE を冗長 PE で置き換えることにより，論理的に欠陥の無い格子結合網を得ている．SRT 網もトーラス網を基本としているため，欠陥回避が容易である [7]．

WSI よりも大規模なシステムを構築する実装法として，WSI を 3 次元的に構築する WSI スタックシステムについても研究が成されている．WSI スタックシステムでは，欠陥回避に加え，内部にあるプロセッサの発生する熱を，どのように放熱するかが大きな問題となってくる．このため，現在試作されている WSI スタックシステムは，かなり小規模のものとなっている．

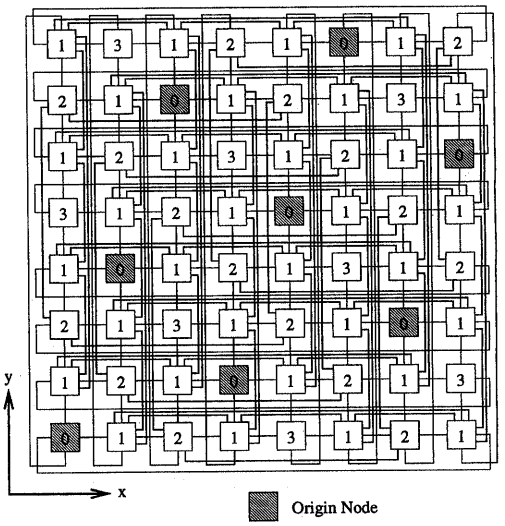

図 1: 8 × 8 ノードから成る 2 次元 SRT 網.

本論文では，科学技術計算に適する結合網である SRT 網の WSI スタック実装について，欠陥回避アルゴリズムを改良することによって WSI スタック内部の最高温度を低下させることを試みる．最初に SRT 網について簡単に述べ，WSI 実装された SRT の欠陥回避方式を示す．欠陥回避の規則に従ってウェーハ内に動作 PE を配置する新たな 2 種類の欠陥回避方式を提案し，WSI スタックシステム全体の歩留まりと内部最高温度について評価を行なう．

2 SRT 相互結合網

2 次元 SRT 網 (2D-SRT) は，トーラス網上のノードにレベル l を割り当て，各ノードは 2^l の長さのバイパスリンクによって遠隔ノードと接続されることにより構成される結合網である．図 1に 2 次元 SRT のレベルと結合の様子を示す．

$N \times N$ ノードから成る 2D-SRT 上で，レベル l のノード (x_l, y_l) は次の式を満すノードとして定義される．

$$\left(x_l - 2^{l-1} + s_x \cdot y_l\right) \ mod \ \min(2^l, 2^T) = 0$$

ここで s_x は x 方向のシフト幅，T は最高レベルの制限値である．ノード (x_l, y_l) は隣接する 4 ノー

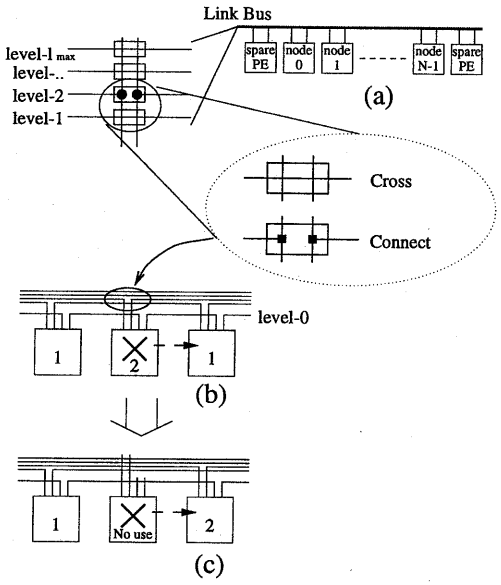

図 2: SRT の 1 次元の故障回避

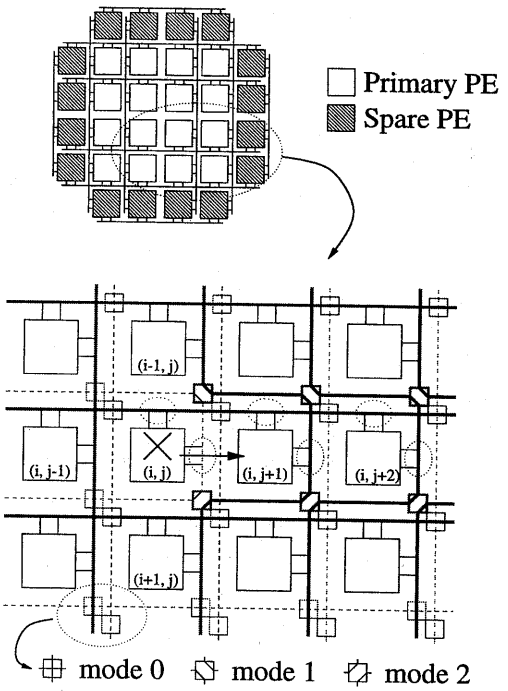

図 3: 2D-SRT の故障回避

ド $((x_l \pm 1) \ mod \ N, (y_l \pm 1) \ mod \ N)$ と，2^l離れた 4 ノード $((x_l \pm 2^l) \ mod \ N, (y_l \pm 2^l) \ mod \ N)$ と接続される．s_x, Tは 2D-SRT の通信性能に大きく影響する．ここでは最も標準的なものとして，$s_x = 2^{\lceil l_{max}/2 \rceil} + 1, T = \log N$とする．

3 2D-SRT の故障回避アーキテクチャ

3.1 1 次元についての故障回避

図 2 (a) に，2D-SRT の水平または垂直方向の，1 次元についての故障回避可能なアーキテクチャを示す．N 個の PE と冗長 PE が直線上に置かれる．各 PE は使用中の状態と未使用の状態の 2 つの状態をとることができる．故障 PE は未使用の状態にされ，隣接する PE によって故障 PE の機能が代替される．代替を繰り返し行うことをシフトと呼び，シフトにより故障 PE を回避する．

シフトを可能とするため，各ノードはレベル 0 からレベル l_{max}までのリンクを束ねたリンクバスに，スイッチ束を介して接続される．各スイッチは Cross モードと Connect モードの 2 つの状態をとることができ，スイッチ束中のスイッチは，0 または 1 つのスイッチだけが Connect モードをとり，他のスイッチは Cross モードにセットされる．このスイッチの切り替えにより，故障 PE の切り離しと代替 PE のレベル切り替えを行なうことができる (図 2 (b) ⇒ (c)).

3.2 2D-SRT の故障回避

図 3に故障回避可能な 2D-SRT のアーキテクチャを示す．$N \times N$のコア PE の周囲に冗長 PE が配置される．PE と PE の間にはリンクバス，故障回避に使用される補償リンクバス，及びリンクバスを切り替える 3 モードスイッチが置かれる．

最初に図に示すような横 (右) 方向のシフトについて考える．横方向シフトはリンクバス上のスイッチ束によって行なわれる．一方，縦方向のシフトはリンクバスは 3 モードスイッチを用いて，リンクバスを曲げることによって行なわれる．

右シフトは 3 ステップの動作により成し遂げられる．ステップ 1 では，故障 PE の縦横方向のリンク

バスに接続する全スイッチ束のスイッチ及びlevel-0スイッチをCrossモードにし，故障PEの切離しを行なう．ステップ2では，補償リンクバス及び3モードスイッチを用いて，縦方向のリンクバスの経路を曲げる．ステップ3では，レベルの再定義を行なう．1D-SRTの故障回避と同様に，ノードのレベルが$l \rightarrow k$に再定義される場合，縦横両方向のリンクバスに接続するスイッチ束のレベルlのスイッチがConnectモードからCrossモードに変更され，代わってレベルkのスイッチがCrossモードからConnectモードにセットされる(図中破線円)．縦方向にシフトする場合も同様の手続きによりシフト可能である．これらの手続きにより，故障PEは上下左右どの方向にもシフトでき，故障の回避が可能である．

上下左右どちらの方向にシフトするかを決定するアルゴリズムは，WSI全体の歩留まりに大きく影響し，いくつかの方法が考えられるが，Kungらはグラフ理論を用いた方法[5]，Numataらはローカル情報のみを用いてヒューリスティックに故障回避を行なう方法(HS法)[6]を提案している．

4 WSIスタック実装

4.1 故障回避と放熱

4.1.1 WSIスタックの発熱モデル

大規模な2D-SRTを構成する場合，複数のウェーハ上に2D-SRTの一部を実装し，このウェーハを層状に重ねることにより，WSIスタックによる実装が考えられる．この概念図を図4に示す．ウェーハ上には実装すべき2D-SRTのサブセットに加え，冗長PEを配置しておく．ウェーハ上の故障PEは，ウェーハ内部で故障回避され，各ウェーハは，論理的に欠陥の無い2D-SRTのサブセットを構成する．これらのウェーハを縦方向に接続し，大規模な2D-SRT網を得ることができる．欠陥回避のためのシフトは，ウェーハ内で完結し，ウェーハ間にまたがるシフトは行わないものとする．

ウェーハ上のPEは，もし動作状態ならば発熱し，非動作状態の場合は発熱しない．放熱をWSIスタックの周囲から行なうとすると，ウェーハ内の発熱量が同じ場合，発熱部分がなるべく周囲に存在する方が，発熱部分が中心部に集中する場合よりも，ウェーハ内の最高温度を低くすることができる．

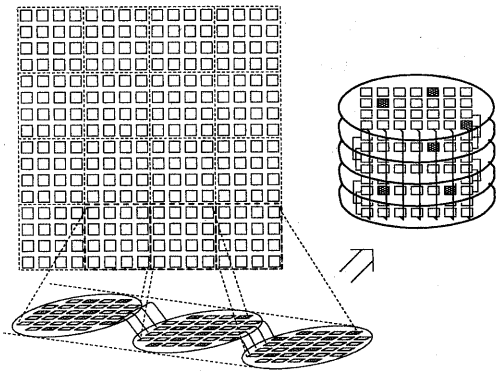

図4: 2D-SRTのWSIスタック実装

欠陥回避の結果，ウェーハ上には動作PE，故障PE，正常だが動作しない休止PEの3状態のPEができる．故障率が十分低い場合，冗長PEの殆どが休止PEとなる．冷却の観点から見ると，故障PEは用いたウェーハにより決定されるが，休止PEは動作PEと交換可能であるため，休止PEをウェーハ中心部に集めることにより，システムの最高温度を低下させることができる．そこで，システムの冷却のために，休止PEをウェーハ中心部に集めるために，冗長PEの集中配置とシフト方向の重み付けの2種類の方式を考える．

4.1.2 冗長PEの集中配置

最初に故障回避を開始する前のPEの初期配置について，冗長PEを周囲に配置する方式と中心部に配置する方式が考えられる．この配置方式の違いについて，システム全体の歩留まりとWSIスタック内の最高温度について考察する．

冗長PEをウェーハの周囲に配置する場合(図5に)は，動作すると仮定しているPEが配置されるウェーハ中心部の表面状態が良く，故障回避の際にどの方向にもシフト可能なので，故障回避後の歩留まりが高いことが期待できる．しかしながら，故障PEが少なくPEの配置が初期配置からあまり変化しない場合，動作PEがウェーハ中心部に集中するため，中心部の温度は高くなる．

冗長PEを中心に配置した場合(図6)は，故障回避の際，シフトの方向が中心方向に限定されるため，ウェーハの歩留りは低下するが，発熱するPE

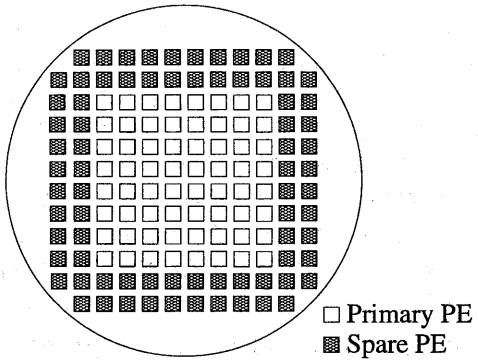

図 5: 冗長 PE を周囲に配置する方式

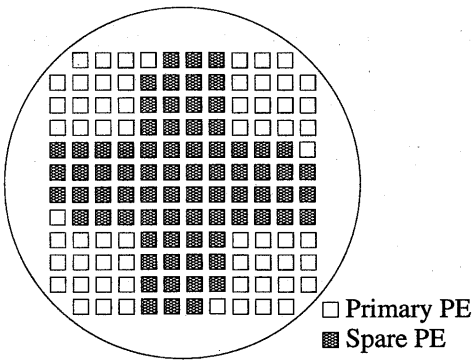

図 6: 冗長 PE を中心部に配置する方式

が周囲に配置されるため，内部温度は低くできる．

4.1.3 シフト方向の重み付け

次にシフト方向の重みづけについて考える．故障回避の際のシフト方向を決定するためには，様々なアルゴリズムが提案されている．HS 法 [6] は非常に高いシステム歩留まりを得ることができるが，シフトの方向は乱数を用いて決定されており，方向は均質である．

そこで，冷却効率を高めるために，重み付けシフトを考える．ウェーハの周囲から放熱するために，動作 PE をできるだけウェーハの周辺に配置されるように，乱数に重み付けを行ないシフトの方向を決定する．この概念図を図 7に示す．シフト方向の重み付けは，x 方向にシフトする確率を p_x とすると，

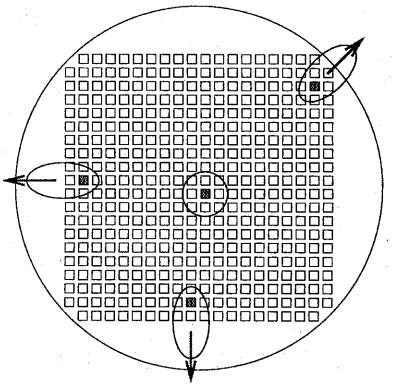

図 7: 重み付けシフト

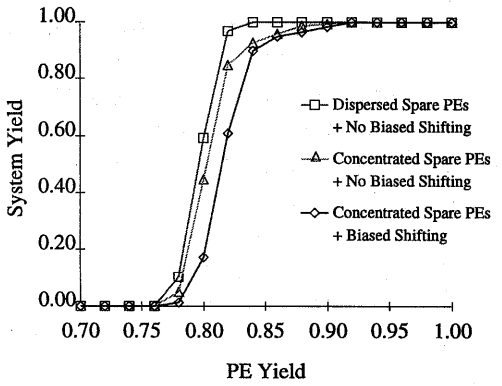

図 8: WSI スタック実装 2D-SRT のシステム歩留まり

シフトの始点となる PE の位置 (x, y) によって，次のように決定される．

$$p_x = \frac{\sqrt{3}}{2}|x| \cdot \left(1 + \frac{|x|}{|x| + |y|}\right)/3 \quad (0 \le x \le 1)$$

4.2 歩留まりとスタック内最高温度

WSI スタックシステムの歩留まりを，(a) 冗長 PE を外周に配置しシフト方向の重み付けが無い場合 (HS 法 [6] と同一)，(b) 冗長 PE を内部に配置しシフト方向の重み付けが無い場合，(c) 冗長 PE を内部に配置しシフト方向の重み付けを行なった場合の 3 つの場合について評価した．

これらの歩留まりを図 8に示す．(a) は最も高いシステム歩留りが得られるのに対し，(c) はかなり

歩留まりが低下する．(b) は (a) よりは低い歩留まりであるが，(a) に比べて著しい低下ではない．(c) の歩留まりが低下する原因は，冗長 PE が内部にあるにもかかわらず，シフトの方向は外向きに重み付けされているため，冗長 PE のある方向にシフトが行なわれにくいためと考えられる．

次に，これらの 3 方式について，ウェーハ内の最高温度をシミュレーションにより求めた．このグラフと，その熱分布の様子を図 9に示す．

シミュレーションの条件として，ウェーハの素材は Si とし，熱伝導率などの物理定数は Si と同じ値を用いた．PE は $(16+4)\times(16+4)$ で構成される．また，1 つの PE の面積を $25mm^2$，ウェーハの直径は $25cm$ とする．周囲温度は 25 ℃，1PE 当りの発熱量を 0.5W とした．

(a) では，PE の歩留まりが 1.0 の時，冗長の休止 PE が全て外周に配置され，内部温度が最も高くなる．故障 PE が増加すると，故障 PE を回避するために，内部の故障 PE が動作せず，外周部の冗長 PE が動作するようになるので，故障 PE が増加するに従い最高温度が低下する．

(b) と (c) では，PE の歩留まりが高いと，外周にある PE が動作し，効率良く冷却を行なうことができる．故障 PE が増加すると，動作 PE が内部の冗長 PE の方にシフトするので，冷却効率が低下する．ウェーハ内最高温度については，(b) と (c) では殆ど差がない．

故障 PE が全くない場合，システム内の最高温度は (a) では約 $410K$ となり，半導体の動作する温度の限界に近くなる．一方，(b) と (c) ではどの歩留まりでも最高温度は $384K$ 以下である．(b) と (c) で，最高温度の差が殆どないことから考えて，PE の配置方式が重要であり，シフト方向の重み付けは温度に関してはあまり影響しない．また，(c) ではシステム歩留まりがかなり低下するので，(b) の場合が歩留まりの低下が少なく，かつ最高温度の低下には有効であることが分る．

5 結論

本論文では，冷却を考慮した SRT の WSI スタック実装について議論した．放熱を効果的に行なうために，欠陥回避用の冗長 PE を中心に配置し，シフト方向の重み付けを行なった．システム全体の歩留まりとスタック内の最高温度をシミュレーションによって求めたところ，冗長 PE を中心に配置する方式が歩留まりを低下させずにスタック内最高温度を大幅に低下させることができることが分った．また，シフト方向の重み付けを行なうと，システムの歩留まりはかなり低下するのに対し，最高温度の低下にはあまり効果が無いことが分った．

今回の方法では，欠陥を回避した後，中心方向に移動可能にもかかわらず周囲に休止 PE が残る場合がある．これら休止 PE の最適な配置が今後の課題である．

謝辞：本研究の一部は文部省科学研究費を用いて行なわれた．関係各位に感謝する．

参考文献

[1] R. Duncan. "A Survey of Parallel Computer Architechtures". *IEEE Computer*, Vol. 23, No. 2, pp. 5–16, 1990.

[2] 楊愚魯, 天野英晴, 柴村英智, 末吉敏則. "超並列計算機に向き結合網:RDT". 信学論, Vol. J78-D-I, No. 2, pp. 118–128, 1995.

[3] W. Worth Kirkman and Donna Quammen. "Packed Expornential Connections — A Hierarchy of 2D-Meshes". In *Proceeding of the Fifth International Parallel Processing Symposium*, pp. 464–470, Apr. 1991.

[4] 井口寧, 堀口進. "超並列計算機向きプロセッサ結合網 SRT". 信学技報, Vol. 95, No. 327, CPSY95-69, pp. 25–30, Oct. 1995.

[5] S. Y. Kung, S. N. Jean, and C. W. Chan. "Fault-Tolerant Array Processors Using Single-Track Switches". *IEEE Trans. on Computers*, Vol. 38, No. 4, Apr. 1989.

[6] I. Numata and S. Horiguchi. "Efficient Reconfiguration Scheme for Mesh-Connected Network: The Recursive Shift Approach". *Proc. of the Parallel Architectures, Algorithms and Networks*, pp. 221–227, June 1996.

[7] 井口寧, 堀口進. "超並列計算機用プロセッサ結合網 SRT — ネットワーク特性と故障回避アーキテクチャ —". 信学技報, Vol. 96, No. 34, CPSY96-45, pp. 37–43, May 1996.

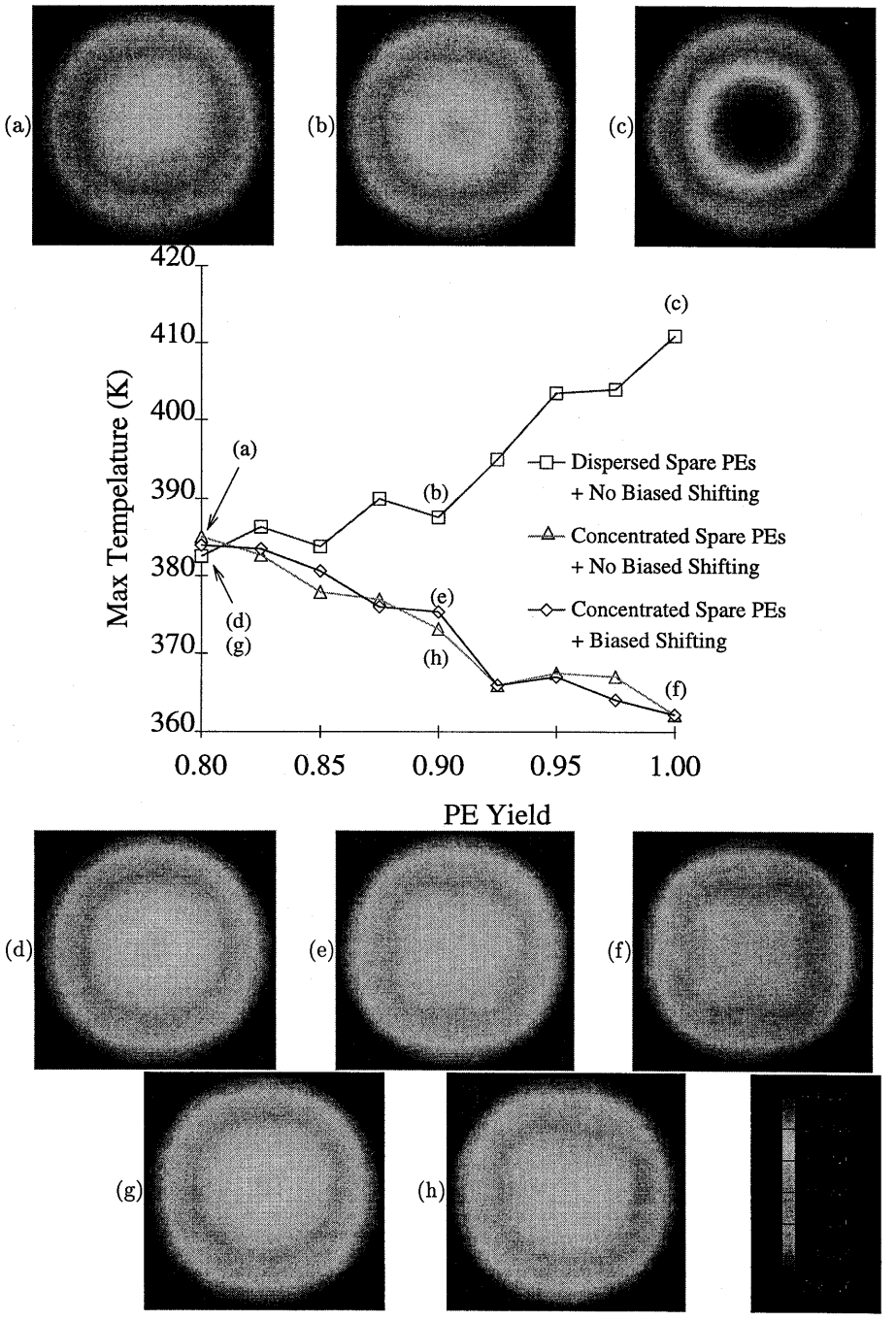

図 9: WSI スタック実装 2D-SRT のスタック内最高温度