マイクロコンピュータとワークステーション 64－4
（1990. 11. 30）

マルチプロセッサ・システムのフォールトトレラント結合方式について

田中 徳彦　　黒川 恭一　　古賀 義亮

防衛大学校 情報工学科

マルチプロセッサ・システムのプロセッサ間結合方式について、パイプライン並列処理を例として相互結合ネットワークとその結合回路（バンクメモリ結合回路）を提案し、フォールトトレラント・パイプライン処理システムを構成する。バンクメモリ結合回路によるマルチプロセッサ・システムは、結合しているプロセッサの機能障害を検出することができるため、フォールトトレラント化に有利な結合回路であるばかりでなく、パイプライン処理のように大量のデータを一方向に転送するアプリケーションに対して従来の結合回路の問題点を改善する。

A Method of Fault Tolerant Communication in a Multi-processor System

Norihiko TANAKA　Takakazu KUROKAWA　Yoshiaki KOGA

Department of Computer Science, The National Defense Academy

This paper proposes a new fault tolerant communication method and three new interconnection networks to construct a multi-processor system for a pipeline processing. The proposed communication scheme using bank memory switching technique has an advantage to make a fault tolerant pipeline system so that it can detect any failure caused in processing elements of the system. In addition, it can overcome the conventional problems caused in interconnection circuits to flow data with one way direction such as a pipeline processing.

1 まえがき

1つの処理を複数の処理ステージに分けて、流れ作業の原理で処理を行なうパイプライン処理は、n台の処理要素（ＰＥ）を用いて処理速度を最大n倍に向上することができる利点を有すると共に、現在までに多数提案されている逐次的アルゴリズムに比較的なじみやすいことなどから、低コストの並列処理機構として多くのシステムが提案され[1]汎用化システムの試みもなされている[2]。しかし、ＰＥを直線状に結合した線形結合マルチプロセッサ・システム上で汎用パイプライン処理を実現する場合、ＰＥや結合回路の故障によって容易にその処理能力が失われる。このためＰＥをネットワーク状に冗長結合して、故障が生じた場合には故障部位を切り離してシステムを再構成し、システム全体の機能を維持するフォールトトレラント化[3]が必要となる。本報告は、パイプライン処理を行なうマルチ・プロセッサシステムのフォールトトレラント化を考慮した結合方式として、2重リング結合、3重リング結合、木状リング結合の3種類の結合方式を提案し、これらの特徴を比較する。2. では、パイプライン処理を行なうマルチプロセッサ・システムのＰＥ間結合について、従来から提案されて来た各方式の問題点を明確にした後、新しい結合方式を提案する。3. では、フォールトトレラント・パイプライン処理のネットワーク結合方式を提案し、3つのネットワークトポロジについてそれらの特徴を比較する。

2 プロセッサの結合方式

2.1 従来の結合方式

パイプライン処理を行なうマルチプロセッサ・システムでは、処理速度の向上を図るために、各ＰＥの仕事量（タスク粒度）を均一にしてＰＥ間でのデータ待ち時間をできる限り少なくする必要がある。このため、非同期のパイプライン処理におけるデータの同期法として、結合回路にＦＩＦＯバッファや2ポートＲＡＭのような記憶装置を用いる方法等が提案されている[1]。このような結合回路は、内部に含まれる記憶装置の形態によって特徴付けることができる。ここでは、この記憶装置として代表的なＦＩＦＯバッファと2ポートＲＡＭを取り上げ、それらの結合回路としての問題点とそのフォールトトレランス性について明確にする。

2.1.1 ＦＩＦＯバッファ

結合回路がＦＩＦＯバッファによって構成されている場合、送信ＰＥがデータをＦＩＦＯバッファに書き込んだ後、今度は受信ＰＥがＦＩＦＯバッファから読み出さなければならないため、ＰＥとＦＩＦＯバッファ間で処理に無関係なデータ転送が2回必要となり、データ転送のオーバーヘッドが大きくなる。また、ＦＩＦＯバッファの容量は比較的小さいので多くのデータを送る場合、ハンドシェイク待ち時間が生じてしまう。

フォールトトレランス性に関しては、ＦＩＦＯバッファは基本的に単方向の通信機能しか有していないため、送信ＰＥはデータを送り出すのみで受信のチェックをすることができない。このため、転送先ＰＥの故障にも関わらずデータを送り続けるような事態が生じる。また、システムの再構成を行なう場合、双方向の通信機能が必要であることから、2つのＦＩＦＯバッファが必要となると共に、再構成を行なうためのメッセージ到着時間がバッファによって不確定になるため、確実な誤り回復が難しくなる。

2.1.2 2ポートRAM

結合回路が2ポートＲＡＭによって構成されている場合においてもＦＩＦＯバッファと同様に、ＰＥと結合回路との間で2回の転送が必要である。しかし、2ポートＲＡＭはＦＩＦＯバッファよりも大きなメモリ容量を実現することができるため、ハンドシェイクの待ち時間はＦＩＦＯバッファの場合よりも減少する。

フォールトトレランス性に関しては、デュアルポートＲＡＭは双方向から自由に読み書きできるため、ソフトウェアによって同期を取ることが容易であり再構成に有利といえるが、逆の面ではハードウェアのサポートがないためにソフトウェアのバク等で隣接ＰＥに障害が波及する可能性がある。例えばデータの受信完了前にデータが新たに書き換えられる等の危険性を常に持っている。

2.2　バンクメモリ結合方式

従来からの結合方式の問題点を改善するために、本報告ではバンクメモリ結合方式を提案する。バンクメモリによる結合方式は、２つのＰＥを結合するために、ＰＥに接続されているメモリブロック自体を交換することによってプロセッサ間でデータを交換する結合方式である。また、ＰＥ同士がスイッチによって分離されているため、ＰＥの機能障害がシステム全体に波及しにくい。更に、結合回路の主要な部分は誤り検出機能を有しており、フォールトトレラントマルチプロセッサ・システムの構成に有利である。

2.2.1　バンクメモリ結合方式の概要

バンクメモリによるＰＥの結合方式を、２ポートＰＥの線形結合網を例として図１に示す。バンクメモリ結合回路は、２つのメモリブロックＭＢ０，ＭＢ１とメモリブロック切り換え回路とから構成されている。プロセッサＰＥ１、ＰＥ２は、メモリブロック切り換え回路によってメモリブロックＭＢ０、ＭＢ１にそれぞれ接続されている。メモリブロック切り換え回路は２つの状態を持っており、図１の例ではＰＥ１と左のＭＢ０、ＰＥ２と２つのＭＢ１、ＰＥ３と右のＭＢ０とを接続している。従って、ＰＥ１が左のＭＢ０に、ＰＥ２が左のＭＢ１にデータを書き込んだ後、そのバンクメモリのスイッチを切り換えるとＭＢ０とＭＢ１が入れ換わり、ＰＥ１とＰＥ２間でデータを交換する双方向リンクが実現できる。

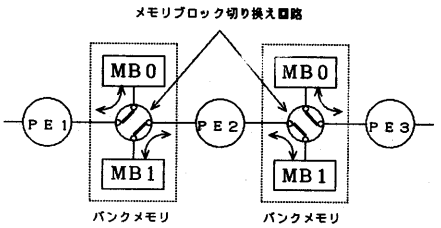


図1　バンクメモリによるＰＥの結合

2.2.2　ＰＥの構成

ＰＥは図２に示すように、ＣＰＵ、ＲＯＭ、ＲＡＭ、Ｉ／Ｏポートとそれらを結び付けるシステムバスとから構成されている。ＰＥのシステムバスからバッファゲートによって分離されているバンクメモリ結合バス（結合バス）には、ＣＰＵの制御により結合バス上でＤＭＡ転送を行うＤＭＡＣ (Direct Memory Access Controller) とバンクメモリとが接続されている。通常動作中バッファゲートは導通状態となっているが、結合バスでのＤＭＡ転送中は、ハイインピーダンス状態となって回路を遮断する。ネットワーク結合方式においては、非隣接プロセッサ間の通信のために、中間に位置するプロセッサの中継が必要であり、一般的には、この中継通信が並列処理の効率を低下させるオーバーヘッドとなっている。この問題点に対して本構成では、バッファーゲートによってＰＥ内バスと分離できる結合バス内でＤＭＡ転送を行うことによって、ＰＥに中継通信のための負荷を負わせずに高速な通信を行なうことができる。

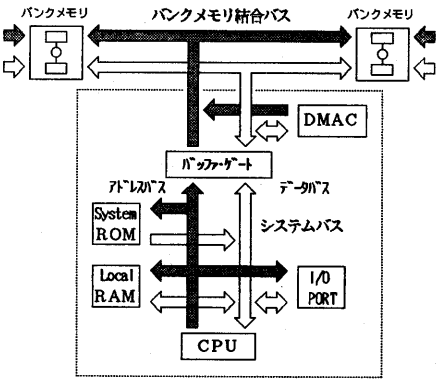


図2　ノードＰＥの構成

2.2.3　多ポートＰＥの構成

ＰＥの多ポート化は、バンクメモリ結合バスにバンクメモリを増設して、各ＣＰＵのメモリ空間内に新たに接続したバンクメモリのための領域を確保することで可能となる。３ポート以上のＰＥの場合、結合バスにブロードキャストモードを付加し、複数のバンクメモリに同時にメッセージを書き込むと共に、ＤＭＡ転送によって１つのバンクメモリから複数のバンクメモリに対して同時にメッセージを転送することが可能となる。３ポートのノードＰＥの構成法を図３に示す。

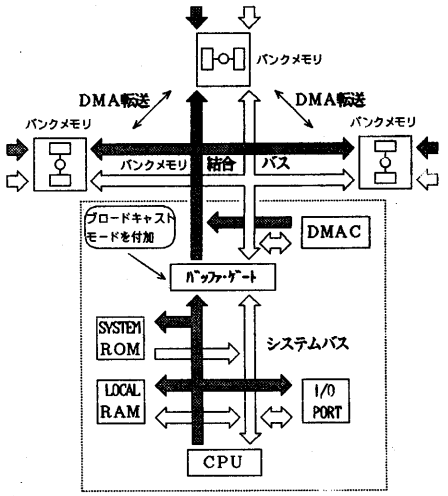
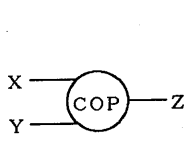


図3　3ポートノードＰＥの構成

2.2.4　メモリブロックの切り換え

メモリブロックの切り換えは、コレスポンデンスオペレータ（ＣＯＰ）とプリセットリミットタイマ（ＰＬＴ）とを用いて、結合されている両方のＰＥの合意のもとに行なわれる。ＣＯＰは双安定出力を持っており、図4に示すように2つの入力の論理が一致したときのみに出力を変化させる。また、ＰＬＴはＰＥ間のデータ転送時間の上限を規定して監視する。転送制限時間からの逸脱は、ＰＥまたは結合回路の障害としてＰＬＴが検出し故障ＰＥを診断する。メモリブロック切り換えの同期機構は図5に示すように3線ハンドシェイクを用いており、双方の切り換え要求とＣＯＰによる切り換え結果とをソフトウェア、ハードウェア双方によって相互に確認しながら確実に行う。



(a) シンボル

INPUTS X	INPUTS Y	OUTPUT Z
0	0	0
0	1	Z_θ
1	0	Z_θ
1	1	1

(b) 真理値表

図4　ＣＯＰ (Correspondence OPerator)

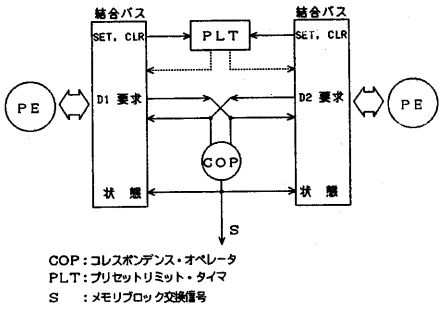


図5　メモリブロック切り換えの合意機構

2.3　従来の結合方式とバンクメモリ結合方式との比較

従来の結合方式では記憶装置が結合プロセッサで共有されているため、データを転送する場合、記憶装置への読み書きの動作が転送速度を低下させる原因となっていた。これに対して、バンクメモリは見かけ上隣接ＰＥで共有されているが、実際はおのおの別個の2つのメモリが独立して入出力に使われているため、結合回路上のデータを直接ＰＥで使用することができる。特にパイプライン処理のように、大量のデータを流しながら処理を行なう場合、結合回路の転送速度の差が問題となりバンクメモリによる結合方式が有利となる。

フォールトトレランス性に関しては、従来の方式では、転送されてきたデータの正当性が時間的にも量的にも厳密に保証されていないのに対して、バンクメモリは入出力のためのメモリが分離されているため、入出力の動作が互いに他に影響することなく安全であり、時間管理の同期機構をも備えていることから、転送データの正当性を保証する事が出来る。

3　フォールトトレラント・パイプライン処理方式

フォールトトレラントシステムを実現するための1つの方法として、システムにあらかじめ冗長な部位を付加しておき、故障が発生した場合はその故障部位を診断して置き換える方法がある。このためには、故障の検出、故障の診断、システムの再構成、誤りからの回復の各処置をオンライン中に自動的に行

なうシステム構成やアルゴリズムが必要となる。

以下では、まず3.1で本報告における議論の対象を明確にするため、ネットワークの前提条件、故障モデル、ネットワークの再構成法の方針をそれぞれ示した後、3.2で3種類のネットワークを順に提案してその特徴を比較する。今回の提案では、3ポートのＰＥによるネットワークを考える。ＰＥのポート数はネットワークの通信距離に関係しており、多いほどＰＥ間の通信距離が短くなるが、あまり多くするとＰＥに対して結合回路が指数関数的に多くなると共に制御ソフトウェアも複雑になり、結合回路の信頼性の低下によるシステム全体の信頼性の低下が問題となる。その中で3ポートは、多様なネットワークを構成することができる最小の次数であり、パイプライン処理のような全順序関係を持つ一次元的な処理には最低2ポートが必要であることから、最小の冗長構成であると言える。

3.1 準備事項

3.1.1 ネットワークの前提条件

- 全てのＰＥは均質で、次数は3（3ポートのＰＥ）であり、バンクメモリによってＰＥ間を直接結合する。
- パイプライン処理はＰＥ数n（nは偶数）のリングネットワーク上で行なう。リング上の0番のＰＥが入出力を担当しており、ＰＥ1から番号順にパイプラインの各ステージが実行されてｎ－1ステージで終了し、出力データをＰＥ0に送る。このリングネットワークを基本リングと呼び、基本リングを構成するＰＥを基本リングＰＥと呼ぶ。
- 冗長プロセッサＲＰを基本リングＰＥの冗長なリンクに結合し冗長ネットワークを構成する。

3.1.2 故障モデル

パイプライン処理を行なうマルチプロセッサ・システムの故障モデルとしては、ＰＥとそのリンクを対象とし、ＰＥが出力を出さず入力も受け付けなくなる沈黙の故障を考える。マルチプロセッサ・システムの一般的な故障モデルでは、故障ＰＥはどのような動作も行ない得るとしているが、本報告で対象とする故障モデルは、ＰＥの入出力に対して厳重なプロトコルを設けることによって妥当性を持つと考えられる。また、このような沈黙の故障の仮定はＰＥの故障によってシステム全体に障害が波及することがなく、隣接ＰＥから容易に故障検出・診断ができる故障モデルでもある。

3.1.3 ネットワークの再構成法

故障が発生した後、システムの再構成を行ない誤りを回復するため、手段としては故障ＰＥを検出・診断して切り離し、予備ＰＥを新たに割り当てることになる。しかし、ＰＥの故障が発生したときパイプライン中にはまだ汚染されていない未処理のデータ列があるため、迅速な誤り回復のためには正常なＰＥは初期の接続状態を維持していた方が、全ＰＥの接続を変更して最適な再構成を行なうよりも効率が良いと考えられる。このため、故障ＰＥのみを代替ＰＥに最短経路で置き換えるものとする。

3.2 結合ネットワーク

ここではリング結合を基礎として、3.1.3で示したネットワークの再構成法に合致する結合ネットワークとして、2重リング結合、3重リング結合、木状リング結合の3種類を提案し、ネットワークの構成方法と再構成アルゴリズムを示す。以下の説明において、P_iはiによって番号付けられたn個の基本リングＰＥの1つであり、正常動作中におけるパイプライン処理の主体である。また、R_iは基本リングＰＥに結合する冗長プロセッサの1つであり、基本リングＰＥの代替プロセッサとする。

［定義1］
プロセッサP_i, R_iの結合を(P_i, R_i)と表現する。

［定義2］
番号付けられたプロセッサ$P_i(i = 0, 1, \ldots, n-1)$が番号順にリング結合したものを$< P_i \mid i = 0, \ldots, n-1 >$と表現する。

［定義3］
$m+1$ビットの2進数nを$(1, n_{m-1}, \ldots, n_1, n_0)$としたとき、置換$\sigma$を次のように定義する。
$\sigma(1, n_{m-1}, \ldots, n_1, n_0) = (0, n_{m-2}, \ldots, n_0, n_{m-1})$
また、逆置換σ^{-1}を次のように定義する。
$\sigma^{-1}(0, n_{m-2}, \ldots, n_0, n_{m-1}) = (1, n_{m-1}, \ldots, n_1, n_0)$

置換σは、nの下位mビットをローテイトシフトし、ＭＳＢをビット反転したものである。

3.2.1　２重リング結合

基本リングＰＥと冗長ＰＥの数を同数とし、基本リングＰＥと冗長ＰＥとで同じ処理を同時に実行しておけば、誤りが検出され次第冗長ＰＥを基本リングＰＥと置き換えることができるので、誤りを速やかに回復することができる。この方式はホットスタンバイ方式と呼ばれており迅速な機能回復ができる。

［構成法］

基本リング$< P_i \mid i = 0,\ldots,n-1 >$と冗長プロセッサによる第２リング$< R_i \mid i = 0,\ldots,n-1 >$の間に$\{(P_i, R_i) \mid i = 0,1,\ldots,n-1\}$の結合を行なう。$n = 8$の時の例を図６に示す。図６の円はプロセッサを表わしており、２重の円は基本リングプロセッサ$P_i(i = 0,1,\ldots,7)$を、また単円のプロセッサは冗長プロセッサ$R_i(i = 0,1,\ldots,7)$をそれぞれ表わしている。円の中の数字は初期動作中そのプロセッサが実行するパイプライン処理のステージ番号である。第０ステージは入出力のタスクであり、パイプライン処理は１から７の７ステージで実行される。

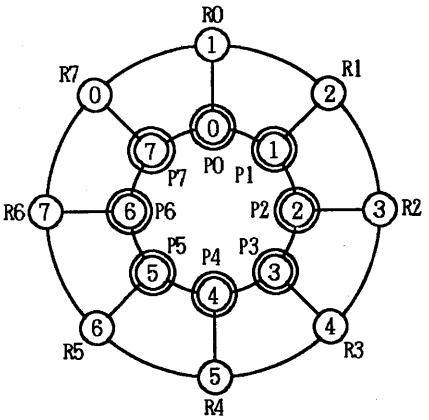


図６　２重リング結合$(n = 8)$

［パイプライン処理］

パイプライン処理は、P_0のプロセッサが入力データをパイプに投入できる状態に準備して、P_1へデータ転送を始めることで開始され、P_{n-1}のプロセッサで最終ステージが終わりP_0プロセッサへ処理結果を送ることで終了する。処理の途中、プロセッサP_iはプロセッサP_{i+1}へデータを送ると共に、プロセッサR_iにも同じデータを送る。プロセッサR_iは常にP_{i+1}プロセッサと同じ処理を行っているが、再構成の指示がない場合はその結果を捨て、新たなデータをP_iから受け付ける。

［再構成アルゴリズム］

基本リングＰＥ中、任意の１つの基本リングプロセッサP_iが故障した場合、故障プロセッサに接続されている前後の基本リングプロセッサP_{i-1}, P_{i+1}と冗長プロセッサR_iが故障を検出する。この位置関係を図７に示す。特に故障プロセッサP_iの前段にあるP_{i-1}はR_{i-1}に対して再構成のため、P_{i+1}との接続を要求し、R_{i-1}が新しいP_iとして処理を補完する。R_{i-1}からP_{i+1}への接続経路は２通りあるが、近い経路（右回りの経路）では２個の中継プロセッサR_i, R_{i+1}が必要であり、遠い経路（左回りの経路）では$n-2$個の中継ＰＥが必要となる。故障の検出から再構成までの時間が短い点から、基本的に中継ＰＥ数の少ない右回り経路を使用する。

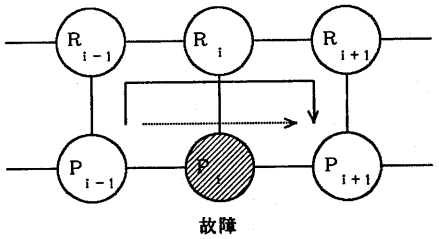


図７　P_iと他のプロセッサの接続関係

［例］

図６においてプロセッサP_2が故障した場合、P_1, P_3, R_2がそれを検出してP_2の代わりにR_1がステージ２の処理を実行し、処理データをP_3に転送する。

3.2.2　３重リング結合

本方式は、基本的には２重リング結合と同様であるが、基本リングＰＥの故障によって冗長ＰＥを置き換える場合、再構成接続のための中継プロセッサ数を減少させるために、偶数番と奇数番の基本リングＰＥにそれぞれ結合する２つのリングを新たに構成するものである。

［構成法］

nが偶数のとき、基本リング$< P_i \mid i = 0,\ldots,n-1 >$と冗長プロセッサによる第２リング$< R_j \mid j =$

$0,\ldots,n-2,j$は偶数 >、第３リング< $R_k \mid k = 1,\ldots,n-1,k$は奇数 >を考え、基本リングと第２、第３リングの間に $\{(P_i,R_i) \mid i = 0,1,\ldots,n-1\}$ の結合を行なう。$n=8$の時の例を図８に示す。

［パイプライン処理］

２重リング結合と同様に基本リング上で処理される。

［再構成アルゴリズム］

基本リングＰＥの１つP_iが故障した場合、冗長プロセッサのR_{i-1}が処理を引き継ぎ、R_{i-1}からP_{i+1}への接続経路が第２リング、第３リングのどちらかのリングによって形成される。２重リング結合と異なり再構成接続のために２つのリングを使用することができるため、右回りの経路では１個の中継プロセッサが必要となり、左回りの経路では$\frac{n}{2}-1$個の中継で経路が構成できる。

［例］

図８においてプロセッサ P_2が故障した場合、P_1,P_3,R_2がそれを検出してP_2の代わりにR_1がステージ２の処理を実行し、処理データをP_3に転送する。

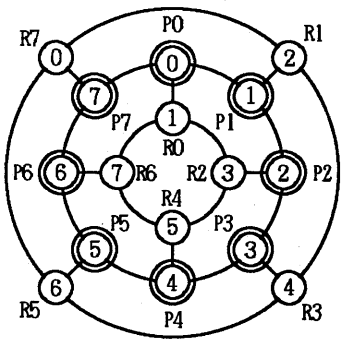


図８　３重リング結合 $(n=8)$

3.2.3　木状リング結合

２重リング結合と３重リング結合を比較すると、３重リング結合の方が再構成接続のための中継ＰＥ数が$\frac{1}{2}$に減少している。木状リング結合は、再構成のための中継ＰＥをなくすために直接代替ＰＥを結合するものである。

［構成法］

レベル数 $m(m=0,1,\ldots,l)$ の２進木の末端の葉ノードの番号 $2^m+i(i=0,1,\ldots,2^m-1)$ をσ置換して、置換後の番号順に基本リングを構成する。$n=8(m=3)$ の例を図９，１０に示す。R_1からR_{15}までの２進木結合網の葉ノードR_8からR_{15}の番号をσ置換してP_0からP_7に置き換え、リング結合させている。

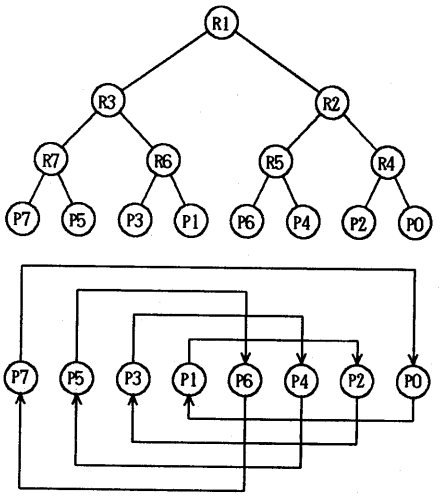


図９　木状リング結合の構成

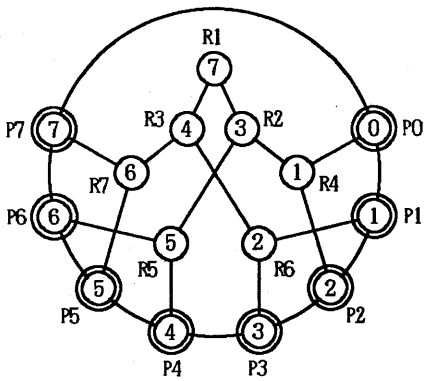


図１０　木状リング結合 $(n=8)$

［パイプライン処理］

２重リング結合と同様に基本リング上で処理される。

［再構成アルゴリズム］

基本リングＰＥの１つP_iが故障した場合、iのσ^{-1}

置換によって再構成経路が決定される。

［例］

図１０においてプロセッサ P_2が故障した場合、P_1, P_3がそれを検出して P_2の代わりに R_6がステージ２の処理を実行し、処理データを P_3に転送する。

3.3 各ネットワークの比較

基本リングＰＥ数を n としたときの冗長ＰＥ数、再構成時の中継ＰＥ数、また、中継ＰＥの信頼度を rとしたときの中継経路の信頼度によって各ネットワークを比較する。

3.3.1 冗長ＰＥ数と中継ＰＥ数による比較

２重リング結合、３重リング結合は冗長ＰＥ数が n であり、中継ＰＥ数も前述のように２個と１個にそれぞれ定まっている。しかし、木状リング結合の場合、再構成のための中継ＰＥ数は故障するプロセッサの位置によって変化する。$n = 2^m (m = 3, 4, \ldots, l)$ のときの木状リング結合の冗長ＰＥ数と中継ＰＥ数は次のように示される。

(1) 冗長ＰＥ数　　$n-1$

(2) 中継ＰＥ数

$2(m-1)-2$ であるものが４通り

$2(m-k)-2$ であるものが 2^k通り

(ただし $k = 2, 3, \ldots, m-1$)

また、これらの中継ＰＥ数を経路数で重み付けをして平均を求めると

(3) 中継ＰＥ数の平均　　$2 - \frac{8}{n}$

3.3.2 中継経路の信頼度による比較

中継ＰＥの１つの信頼度を rとしたとき、２重リング結合の中継経路の信頼度 $R_{2c}(r)$ と、３重リング結合の中継経路の信頼度 $R_{3c}(r)$ は中継ＰＥの直列結合と見ることができるから次のように示される。

$$R_{2c}(r) = r^2$$

$$R_{3c}(r) = r$$

木状リング結合では、再構成の経路によって中継ＰＥ数が違うことからそれぞれの経路毎に信頼度を求めてその平均を取る。平均の信頼度 $R_{tc}(r)$ は

$$R_{tc}(r) = \frac{1}{n}\left(4r^{(2m-4)} + \sum_{k=3}^{m} 2^{k-1} r^{2(m-k)}\right)$$

$n = 8$ の場合の平均の信頼度 $R_{tc}(r)_{(n=8)}$ は

$$R_{tc}(r)_{(n=8)} = \frac{1}{2}r^2 + \frac{1}{2}$$

となり、これは２重リング結合、３重リング結合の信頼度よりも高い。

表１　各ネットワークの比較

比較項目	２重リング結合	３重リング結合	木状リング結合
基本ＰＥ数	n	n	n
冗長ＰＥ数	n	n	n-1
中継ＰＥ数	2	1	2-8/n
信頼度(n=8)	r^2	r	$r^2/2+1/2$

4 あとがき

汎用パイプライン処理のためのフォールトトレラント結合方式として２重リング結合、３重リング結合、木状リング結合の３種類を考え、それらの評価を行なった。バンクメモリ結合回路を用いたマルチプロセッサ・システムは、大量のデータをプロセッサ間で転送するパイプライン処理において、従来の結合回路よりも有利であると共に、故障プロセッサを検出して再構成を迅速に行なうことができる。また、各ネットワークは、ホットスタンバイ方式の冗長構成であるが再構成のアルゴリズムと再構成接続経路の信頼度等にそれぞれ特徴がある。その中で３重リング結合は、再構成アルゴリズムが簡単であると共に、再構成時の接続経路が短く有利であると言える。

参考文献

[1] 高橋義造 編：並列処理機構, 丸善 (1989).

[2] 遠藤, 中村, 重井：汎用パイプライン処理における機能割付の最適化, 信学技報, EC82-38(1982).

[3] 南谷 崇： 並列処理におけるフォールトトレランス技術, 情報処理, Vol.27, No.9, pp.1039-1048(Sep. 1986).