システムソフトウェアと
オペレーティング・システム 69−15
（１９９５．６．２３）

RWC−1 相互結合網用プロトタイプ・ルータの設計

横田 隆史　松岡 浩司　岡本 一晃　廣野 英雄
堀　敦史　坂井 修一

新情報処理開発機構 つくば研究センタ

我々はすでに超並列計算機 RWC−1 向け相互結合網として、小次数で高い転送性能を得られるなど超並列向けに好ましい特性を持つ新しい直接のクラス MDCE (Multidimensional Directed Cycles Ensemble extension) を提案している。本稿では、十分な性能を得るために MDCE 網を採用し、さらに効率的な運用を可能にするために空間分割・時間分割などの運用支援機能を取り入れた相互結合網ルータについて検討するとともに、超並列システム構築へのステップとして試作したプロトタイプ VLSI の設計内容について報告する。ルータ機能は 3 チップスライスで実現される。チップは 0.6μm ルール CMOS ゲートアレイで試作され、約 17 万ゲートの規模を持つ。

A Prototype Router VLSI for Massively Parallel Computer RWC−1

Takashi Yokota　Hiroshi Matsuoka　Kazuaki Okamoto　Hideo Hirono
Atsushi Hori　Shuichi Sakai

Tsukuba Research Center, Real World Computing Partnership

We have proposed a low-degree and high-performance interconnection network MDCE (Multidimensional Directed Cycles Ensemble extension) for the massively parallel computer RWC−1. This paper describes the first router VLSI chip realizing three dimensional MDCE networks. We first discuss required functions for effective system management such as partitioning and time-sharing, and many of them are implemented in the router. We then explain the outline of the prototype chip design. A router is realized by cooperating 3 identical chips. The chip has about 174,000 gates and has been designed and fabricated in a 0.6μm CMOS gate array.

1 はじめに

RWC-1はリアルワールド・コンピューティング(RWC)研究計画の一環として研究開発が進められている1,000台規模の超並列計算機プロトタイプである。RWCで行なわれる広範な応用領域をカバーするのみならず、超並列アーキテクチャの研究上さまざまな試みが行なわれている。

RWC-1はRICA (Reduced Inter-processor Communication Architecture, [5])と呼ばれる方式をノードアーキテクチャとして採用している。これはマルチスレッド・アーキテクチャをさらに発展させた方式であり、RISCベースのプロセッサ・コアにメッセージハンドリング機構を直接的に融合した形式をとる。RICAによれば、プロセッサ間通信のメッセージ(以降パケットと称する)はレジスタファイルから直接生成され、即座に相互結合網に投入される。パケットは目的のノードに到着後、直ちにレジスタファイルに注入され、同時に起動される対応スレッドにより即座に消費される。RICAの導入によりパケットハンドリングのコストが大幅に削減され、効果的な超並列処理が期待されている。

一方、RICAのより効果的な実現のため相互結合網には、高スループットと低レイテンシ、超並列での現実的な実現コスト、といった厳しい要件が課せられる。これに対し、我々はすでに新しい相互結合網トポロジMDCE (Multi-dimensional Directed Cycles Ensemble extension)を提案し、さらに超並列OSへの対応なども含めたRWC-1への実装の指針を示している[7, 8]。

本稿では、上記指針に基づいて設計・試作されたプロトタイプ・ルータチップについて報告する。なお、以下本稿では、RWC-1プロセッサ・チップをPE (Processing Element)、ルータ・チップをSU (Switching Unit)と略称する。

2 MDCE (Multidimensional Directed Cycles Ensemble)網

2.1 DCE (Directed Cycles Ensemble)網

最小次数[1]の結合網は単方向リングである。簡単であるが小規模システムならば十分に適用可能な特性が得られる。この単方向リングをもとに、各ノードで入・出次数を各々1だけ増して複数のリングを相互に結ぶことを考える。これによってできる結合網のクラスをDirected Cycles Ensemble (DCE)と称している。DCEは有向連結グラフとなる。

DCE網は比較的広い範囲の直接網をカバーする結合網クラスである。例えば電総研EM-4, EM-Xで採用されているサーキュラ・オメガ網もDCE網のバリエーションのひとつである。この網は[4]等で報告されているように、低次数でかつ良好な転送特性が得られる。

本稿では典型的なDCE網の例として以下のものを考える。n個のノードで構成される単方向リングを2^n並べる。リング内でのノード位置をx軸成分で表現し、そのノードが属しているリングの位置をy軸成分で表現すれば、ノードのアドレスは(x,y), $0\leq x<n$, $0\leq y<2^n$と表現できる。各ノードでは出次数2のうち一方を単方向リングの形成に使用し、他方をリング間の接続に使う。前者を並行リンク(parallel link)、後者をクロスリンク(cross link)と呼ぶ。各々の接続先は、上記の(x,y)アドレスを使い、

$$(x,y)\rightarrow\begin{cases}((x+1)\bmod n,\ y) & \text{(平行リンク)}\\ ((x+\delta)\bmod n,\ f(x,y)) & \text{(クロスリンク)}\end{cases}$$

と表現される。$a \bmod b$はaをbで割った剰余、δは非負の自然数、$f(x,y)$はクロスリンクでの接続関係を規定するマップ関数を示す。

本稿では、δ=1, $f(x,y)=y\hat{}(1\ll x)$で表現されるサーキュラ・バンヤン網(図1参照。以下c-Banyanと略称)、および、上記c-Banyanの表現からδ=0とすることによって得られるCCC (Cube-Connected Cycles [3])[2]の2つの網をDCE網クラスを代表するものとして扱う。ここで演算記号$\hat{}$,$\ll$は、各々ビット毎の排他的論理和、左シフトを表している。

2.2 DCE網の多次元拡張

さらに超並列システムに対応するため、直積によってDCE網を多次元に拡張することを考える。たとえば2次元平面上で表現された2つのDCE網の直積は図2で示される。

図2中、x軸方向に走っている平行リンクは直積をなしている2つのDCE網で共通であり、したがって各ノードの次数をさらに1増すだけで1次元の拡張が可能である[3]。このようにして構成される網を我々はMDCE (Multidimensional DCE extension)と呼んでいる。MDCE網は任意のDCE網の直積として定義される。

特に典型的なMDCE網として、B次元のc-BanyanとC次元のCCCの直積で求められるものを、並行リンクの多重度Pを加えて(B,C,P)-MDCEと表現する。

[1] 本稿では、ハードウェア実現上の観点から、結合網トポロジを有向グラフとしてとらえ、次数を入次数・出次数の和として論ずる。

[2] リング部分が単方向のため正確にはサブセットとなる。

[3] ただし隘路になるのを防ぐために多次元MDCE網では平行リンクを多重化する必要がある。

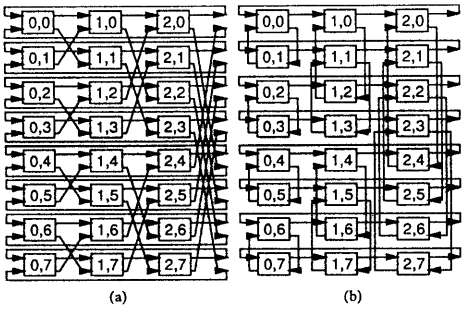

図 1: DCE 表現によるサーキュラ・バンヤン網 (a) および CCC 網 (b)

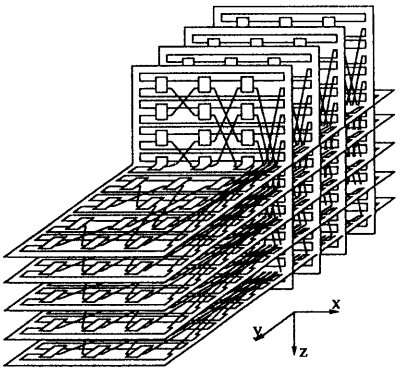

図 2: DCE 網の直積

3 RWC–1 ルータ・チップが提供する機能

3.1 トポロジ

RWC–1 用の相互結合網には (1, 1, 1)-MDCE 網の採用が検討されている [7]。このトポロジは c-Banyan 網を多次元立方の頂点に配置して相互に接続したものとの説明も可能であり、CCCB (Cube-Connected Circular Banyans) とも呼ばれる。以降本論文では CCCB の呼称を用いる。

1,024 ノードの CCCB 網は、3 次元空間内に $x \times y \times z = 4 \times 16 \times 16$ 個のノードを配置し、ノード (x, y, z) から

$((x+1) \bmod 4,\ y,\ z)$	(平行リンク)
$((x+1) \bmod 4,\ (y \hat{}\ (1 \ll x)),\ z)$	(クロスリンク (y))
$(x,\ y,\ (z \hat{}\ (1 \ll x)))$	(クロスリンク (z))

の 3 ノードに接続することで構築される。上記接続規則に従って適切なルーティングを行なう機能が RWC–1 ルータチップの第一要件である [7]。

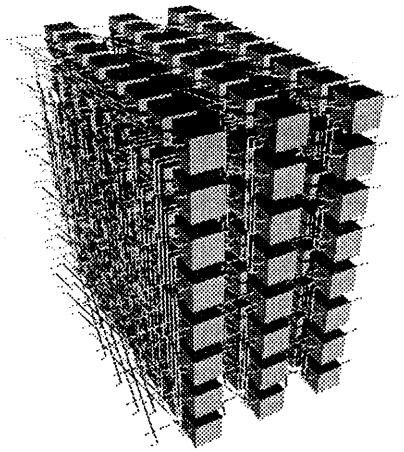

図 3: CCCB 網の例

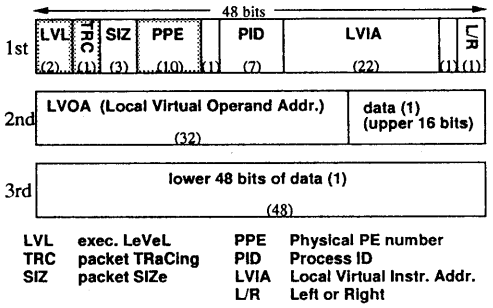

図 4: パケット形式

3.2 パケット形式

本ルータでサポートされているパケットの形式を図 4に示す。パケットは 48 ビットを転送単位 (パケット語) として、PE および SU の内部動作クロックと同一のレートで転送される。

先頭および 2 番目の語の一部はヘッダに相当する。命令ポインタ + データフレームポインタの組をベースとした continuation に、パケット処理上必要な諸情報を加えた形式となっている。ヘッダは第 2 ワードの途中で終り、以降、転送の対象となるデータが上位ビットから順に詰め込まれる。PE での処理単位およびデータの転送単位は 64 ビット固定であり、48 ビット幅のパケットと整合するために PE 内部に変換・緩衝回路を持つ。ひとつのパケットで転送されるデータは 1 ～ 8 個に制限されている。

ルータが参照するのは先頭語に含まれる LVL, TRC, PPE 領域のみである。ルータ内部では 2 ワード目以

降は単なるペイロードとして扱われる。

3.3 転送制御機能

バーチャルチャネル バーチャルチャネルはStore &Forwardデッドロックを防止できるだけ実装する必要がある。MDCE網に対しては、サーキュラ・オメガ網で提案された螺旋バッファ法[4]が適用可能である。CCCB網での必要チャネル数は3である。

フロー制御 RWC-1ではパケットの粒度が小さいことから、1個のパケット全体をルータ内部のバッファに収めることが可能である。パケットの転送は、通常ルータ間でパイプライン的に転送し、ブロックされた時のみルータ内のバッファ内に保管するvirtual cut-through方式を採用する。

ハンドシェーク制御はパケットの転送開始時のみに行なわれる。一度転送が開始されたパケットは次のルータ(あるいはプロセッサ)に向かって、途中で止められることなく最後まで連続して送られる。

優先度制御 RWC-1のパケットは4つの実行レベルを持つ。上位2レベルを`system`, 下位を`user`とし、各々別個のバッファを割り当てる。常に`system`側が高優先度を持ち、`user`パケットは`system`バッファが空の状態の時にのみ転送される。

プロセッサ入力バッファ(*PIB*) 目的の配送先に到着したパケットは、4つの実行レベルに応じて設けられている別個のバッファに一時保管される。個々のバッファは、デッドロックを回避するために見かけの容量を無限大とする必要があるため、チップ内部に設けた容量を越えた分を蓄えるためのバッファ領域(Auxiliary Packet Buffer, 略称*APB*)をローカルメモリ上に確保し、ソフトウェアの介在なしに一時退避/回復する機能を実現している。

3.4 OS支援機能

空間分割支援機能 MDCE網では、ノードの3次元座標アドレスのy, z軸成分のビットパターンにより2の累乗個ノードだけ割り当てるように分割すれば、パーティション間での通信の干渉がない閉パーティション[1]を実現できる。ルータはパーティション制御のための専用のレジスタを備え、レジスタの設定内容に反してパーティション外に出ようとするユーザ・パケットをエラー処理する。ただしシステムパケットは、複数パーティションに渡っての制御を行なう場合があるため、ここでのパーティション管理の対象としていない。

時間分割支援機能 並列処理環境では結合網もプロセススイッチ時の資源管理の対象である。RWC-1結合網では、アプリケーションおよびOSからの要請により、ルータ上に残っている未配送パケットを強制的にローカルメモリ内に保存・復帰する機能(ドレインと呼ばれる)を実現している。

ルータは、ドレインによるOSの時間分割管理を支援するため、次の2つの機能が実装される。

ドレイン同期伝搬機能 未配送パケットの保存・回復を行なうには、(パーティション内)全ノードの同期を取り、ルータでのパケット配送を凍結しておく必要がある。このために必要な同期操作を最小オーバヘッドで実現するための補助機能である。直接接続されているルータ間で局所的に同期が完了するとPEに対して割り込みを発生する機能を持つ。

バッファ内容の排出/回復機能 SUチップ内に残った未配送パケットの保存、前回保存パケットの回復は、SU, PE間でDMA転送機能を実現することで自動的に行なわれる。転送の完了は上記同様、割り込みで通知される。

3.5 その他の機能

チップ内の制御および状態モニタの機能 ルータには簡単なコマンド体系(SUコマンド)が実装される。PEは、ルータに対して必要なコマンドを送り、内部レジスタの設定や状態のモニタを行なう。

ノードレベル・デバグ支援機能 ハードウェアの動作検証や低レベルソフトウェアのデバグを支援するため、ノードごとに1クロックずつステップ実行できる機能を備える。チップ内部動作を停止した状態でも、内部レジスタの読み出しや再設定が可能である。

4 SUチップ・プロトタイプの実現

4.1 ノード構成

我々は超並列計算機構築への最初のステップとして、機能検証用の小規模実験システム(テストベッド)を試作している。このシステムでは、(a) RWC-1用に開発したVLSI (PE[2], SU)の機能検証、(b) RWC-1低位ソフトウェアの開発、を主な目的としている。ノード構成の概略を図5に示す。動作の確実さおよび扱い易さの点から、SU-SU間の接続にはTTLレベルの同期転送を用いている。ただし、スループットを落すことなく動作周波数を抑えるために、ビット幅を倍(96ビット)にして、転送周波数をチップ内部クロックの1/2にしている。

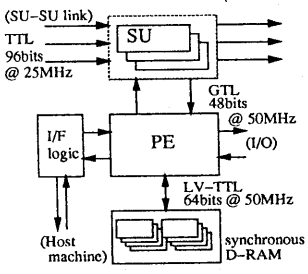

図 5: ノード構成

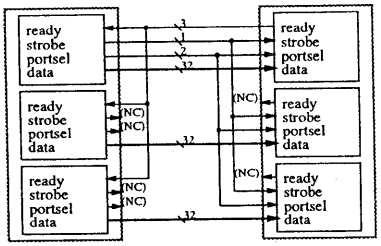

図 6: SU 間の相互接続

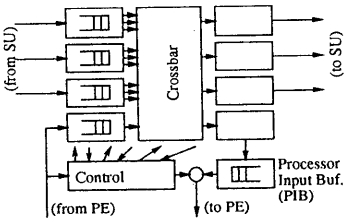

図 7: チップの全体ブロック図

4.2 チップ分割

LSI ピン数制限の問題からルータ機能を 1 チップで実装することはできない。このため本 SU チップでは、96 ビットの物理転送幅を 32 ビットずつ 3 チップにスライスして実装している。各チップは完全に同期して動作し、全体としてひとつのルータ機能を提供する。ルータ相互の接続の様子を図 6に示す。

こうしたスライス化によってパケットの配送制御に必要な情報が全チップに分配されない問題が生じる。そこでパケット形式上、先頭語の最上位 16 ビット内にルータでの配送制御に必要な LVL, TRC, PPE の情報を配し、この部分を担当するスライスをマスタ、他をスレーブとした。マスタチップにはルーティングのための完全な情報が送られる。

ここでマスタからスレーブへの制御情報の伝搬が問題となる。マスタは、次段ルータでのルーティング演算、すなわち、転送中のパケットは次段ルータのどのポートから出力すべきかを計算し、2 ビットの転送制御信号 (PS, PortSelect) としてパケット本体と同時に次段ルータに伝送する (図 6参照)。ルーティング演算をパケット入力後に当該ルータで行なう必要がないため、配送制御の時間的ロスを圧縮できる。

4.3 SU チップの内部構成

SU のブロックダイアグラムを図 7に、SU–SU 間の入力ポートの構成を図 8に、出力ポートの構成を図 9に各々示す。

前段 SU から送られてきたパケットは、まず入力ポート入口でビット幅変換される。チップあたり 32 ビット /($\frac{1}{2}$ クロック) で送られてきたパケットを、16 ビット /($\frac{1}{1}$ クロック) に変換する (図 8)。出力ポートでは、これと逆の操作が行なわれる (図 9)。

入力ポートにはバーチャルチャネル (VC) 数に対応した 3 組のバッファが用意されている。各バッファはさらに実行レベルによってユーザ用 / システム用に分割されているが、優先制御が行なわれ両者同時に動作することがないことから、見かけ上 3 つの FIFO に見える。各バッファの容量は 64 ワードである。PE からの入力ポートは VC をサポートする必要がないため、1 組のユーザ／システム用 FIFO バッファがあるのみである。このバッファも上と同様に 64 ワードの容量である。

これらの FIFO バッファは VC 毎に独立して動作するために、出力部への切替を行なうクロスバスイッチは 10 入力 4 出力となる。クロスバスイッチの機能は、各出力ポートにアービタとマルチプレクサを持たせることで実現している。VC 間の調停は、前回勝ったチャネルの優先度が最低になるように、ラウンドロビン方式で行なわれる。

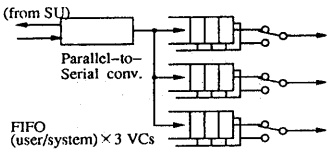

図 8: SU 入力部の構成

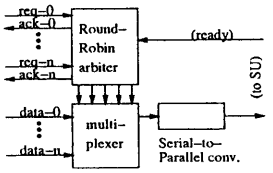

図 9: SU 出力部の構成

4.4 SU チップの実装

本 SU チップの実装設計は、すべてハードウェア記述言語 Verilog-HDL[6] を用いて、RTL (*Register Transfer Level*) で行なった。記述ソースは、総計約

13,000行である(ただし、若干のテスト用記述やコメント等を含む)。

入出力PAD部を除き、コア部のすべてのランダムゲートは、Verilog-HDLによるRTL記述をもとに、市販の論理合成ツールによって生成した。PAD部もVerilog-HDLでネットリストを表現しているため、すべての設計を通して回路図面は全く作成していない。

SUチップは、プロセスルール0.6μmのCMOSゲートアレイで試作されている。パッケージは447ピンCPGAである。表1に主な諸元を示す。

表 1: SUプロトタイプ・チップの主な諸元

Fabrication:	
technology	0.6μm CMOS gate array, 2 metal layers
chip size	15.2 mm × 15.2mm
package	447-pin ceramic pin grid array
I/O signals:	
GTL	22 inputs, 22 outputs
TTL	133 inputs, 129 outputs
direct	3 inputs (GTL ref. voltage)
(total)	158 inputs, 151 outputs
Gate count:	
RAM	59,722 gates, (128w×19b) × 10
	61,501 gates, (1024w×17b) × 1
random logic	52,672 gates

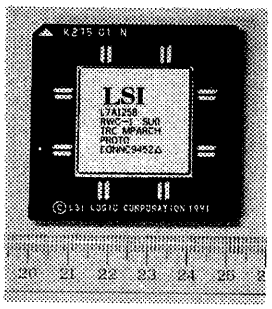

図 10: 試作チップの外観

5 おわりに

本稿では、超並列計算機RWC-1での相互結合網用として研究開発されているルータチップのプロトタイプ試作について報告した。本プロトタイプ・ルータチップは、超並列向きに提案されているMDCE (Multidimensional Directed Cycles Ensemble extension)網を実現するほか、見かけ上無限大のバッファを実現するAPB (*Auxiliary Packet Buffer*)機能、ルーティング情報を前段ルータで演算する機能などの新しい機能が導入されている。また、本ルータチップは、こうしたパケット転送機能と同時に、システムの効率的な運用のために要求されているプロセス切替え時のパケット排出機能、パーティション制御機能を備えている。

プロトタイプ・チップは、0.6μmルールのCMOSゲートアレイで試作されている。回路規模は約174Kゲートである。

今後、開発したルータチップについて、テストベッド・システム上で動作検証を進めながら、本チップで採用されている転送機能およびシステム運用支援機能の効果・有効性を検証する。また、伝送方式や実装方法など今回の実現で残された問題を検討・解決し、1,000ノード規模の超並列システムを構築する予定である。

謝辞

本研究の機会を与えていただいたRWCつくば研究センタ島田潤一所長、MDCEの原形となるヒントを頂いた慶應義塾大学天野英晴助教授、また、有益な議論を頂いたRWC関係各位に深く感謝します。

参考文献

[1] 堀敦史ほか. 並列計算機オペレーティングシステムカーネルSCoreにおけるプロセス管理とハードウェア支援機能. コンピュータシステム・シンポジウム論文集, pp. 59–66, 10月 1993.

[2] 松岡浩司ほか. 超並列計算機RWC–1用プロセッサチップの設計. 信学技報, CPSY95-18, pp. 55–62, 4月 1995.

[3] F. P. Preparata and J. Vuillemin. The cube-connected cycles: A versatile network for parallel computation. *Communications of the ACM*, Vol. 24, pp. 300–309, May 1981.

[4] S. Sakai, et al. Design and Implementation of a Circular Omega Network in the EM-4. *Parallel Computing*, Vol. 19, No. 2, Feb. 1993.

[5] S. Sakai, et al. Reduced Interprocessor-Communication Architecture for Supporting Programming Models. In *Proc. Conf. on Massively Parallel Programming Models*, pp. 134–143, 1993.

[6] E. Sternheim, et al. *Digital Design and Synthesis with Verilog HDL*. Automata Publishing, 1993.

[7] 横田隆史ほか. 超並列計算機RWC–1の相互結合網. 情処学アーキテクチャ研究会, ARC–101–4, 1993.

[8] 横田隆史ほか. 超並列向け相互結合網MDCEの提案と評価. 並列処理シンポジウムJSPP '94, pp. 121–128, 1994.