計算機アーキテクチャ 113−11
（1995. 8. 24）

RWC−1の階層型MDCE網

横田 隆史　松岡 浩司　岡本 一晃　廣野 英雄　坂井 修一

新情報処理開発機構　つくば研究センタ

我々がRWC−1向けに提案している新しい直接網のクラスMDCE (Multidimensional Directed Cycles Ensemble extension)は、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が、プリント基板から引き出せる信号数など実装面で物理的制約を受けやすく、改善の余地がある。本稿ではMDCE網が持つ優れた転送特性を活かしつつ実装性を高めるために、MDCE網の各ノードを下位のDCE網で置き換える形式の階層型MDCE網(c-MDCE)を提案する。c-MDCE網の採用によって、VLSI, プリント基板、筐体の各レベルでの実装性が大幅に改善される。階層化のために転送特性が若干悪化するがなおメッシュ等と比較して優位性を保っている。

The Clustered Multidimensional Directed Cycles Ensemble Networks for the RWC−1

Takashi Yokota　Hiroshi Matsuoka　Kazuaki Okamoto
Hideo Hirono　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 network requires large PC board intersection width to attain sufficient transmission bandwidth. This causes implementation problem in organizing massively parallel RWC−1. In this paper we propose a new MDCE-based network that has the same topology with existing MDCE network with each node replaced by a four-node DCE network. This network, called clustered MDCE (c-MDCE), is a solution for VLSI-, PC board-, and rack-level pin limitation problem. The network property is slightly degraded but it is still advantageous.

1 はじめに

RWC-1[5] はリアルワールド・コンピューティング (RWC) プロジェクト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 で行われる広範な応用領域に適合する超並列処理環境の提供を目的としつつ、同時に、超並列アーキテクチャの研究上さまざまな試みがなされている。

超並列処理環境の実現のためには、基本となる並列処理のモデルとそれを効率良くサポートするための要素プロセッサのアーキテクチャ、さらにスケーラブルな相互結合方式の検討が必須である。RWC-1 ではマルチスレッド・アーキテクチャを超並列向けにさらに発展させ、局所処理, 通信インタフェイス, 同期操作の高速化のために、RICA (Reduced Inter-processor Communication Architecture, [4]) と呼ばれる方式を提案し、ノードアーキテクチャとして採用している。RICA は 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。プロセッサ間通信のメッセージ (パケット) 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の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。

RICA の採用により細粒度処理の可能性が拡がるが、これを超並列に適用するためには低レイテンシかつ高スループットなど基本特性に優れた相互結合網が求められる。さらに RWC-1 実現のためには、ハードウェアコスト (超並列での実現性)、OS 支援機能 (効率的な運用) など、同時に満たすべき要件がある。この要件をよく満たすものとして、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、それを実現するためのルータチップを試作してきた [7, 6]。

超並列計算機の実現に際しては、方式自体の特性とは別に実装性が問題となる場合が多い。特に超並列計算機での相互結合網は物理的制約を考慮する必要があり、多くのマシンで懸案事項となる。MDCE 網によれば、低次数のルータで直径が小さく高スループットの相互結合網を構築できるが、1 ボード上に実装されるノード数が多くなるとボードの次数が多くなる。さらに RWC-1 では転送バンド幅を確保するために多ビット同時転送を行なうため、ボード間を接続する信号数が多くなり実装性が悪化する。

このため MDCE をさらに超並列向きにするために、トポロジを階層化することによってボード次数を低く抑えることを検討した。以下本稿では、MDCE 網の概略を紹介したのち、その転送特性を継承しつつ実装性を大幅に改善した新しい相互結合網階層型 MDCE 網を提案し、その可能性について検討する。

2 MDCE 網

2.1 DCE 網

RWC-1 には広範な RWC 応用領域に適用できる汎用性が求められ、またメッシュなどトポロジ上の制約がないことから、低次数、短直径、高スループットといった基本特性に優れた直接網を模索することから始まった。

最小次数[1]で実現される結合網は単方向リングである。非常に簡単な構造であるが、小規模システムであれば十分な性能を得ることができる。

この単方向リングを第 1 ステップとして、これを拡張することを考える。単方向リン

[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。

グの各ノードに入力・出力ポートを各々1組加え、複数のリングを相互に結合する。これによってできる結合網のクラスをDirected Cycles Ensemble (DCE) と呼んでいる。

典型的なDCE網は次のように構成される。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、

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

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、$\delta = 1$, $f(x,y) = y\hat{}(1 << x)$ 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 $\delta = 0$ 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$\hat{}$, $<<$ 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

DCE網をベースに、さらに超並列システムに対応するため、直積によって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のように求められる。

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1増す

[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。

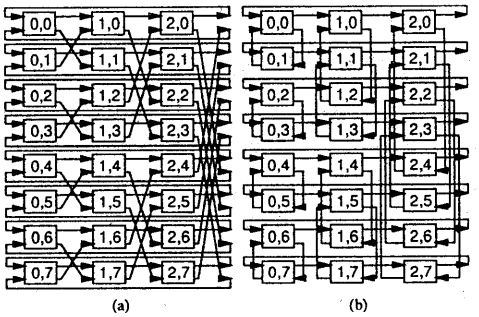

図 1: DCE表現によるサーキュラ・バンヤン網 (a) およびCCC網 (b)

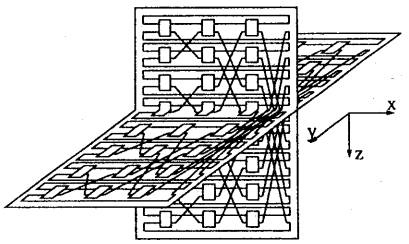

図 2: DCE網の直積

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。たとえば $(1,1,1)$-MDCE の場合、次数は3+3 $(= B+C+P)$ であり、ノード (x,y,z) の3つの出力ポートは

$$\begin{aligned} &(((x+1) \bmod n), y, z) \\ &(((x+1) \bmod n), (y\hat{}(1 << x)), z) \\ &(x, y, (z\hat{}(1 << x))) \end{aligned}$$

の3つのノードの入力に接続される。

[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。

3 階層型MDCE網

3.1 MDCE網の階層化

相互結合網の方式および実現方法の選択は、VLSIのピン数、ボード間の接続に必要な信号線数などの物理的制約を満たしながら、転送性能と実現コストのトレードオフによって決まる。超並列ではできるだけ実装密度を上げる必要があるため、1ボード上に多くのノードが実装される。一般的に1ボードに実装されるノード数が多くなるほど、ボード間接続のために必要な信号線数が多くなるため、実装技術から来る物理的制約およびコストの問題が大きくなる。

MDCE網では、個々のノードの次数は小さく抑えられているものの、複数ノードを実装した時のボードの入出力ポート数(以降ボード次数と呼ぶ)を小さく抑えられず、ボード間の接続性が問題となる。たとえば(1,1,1)-MDCE網では、図3に示すように1ボード上に4ノードを実装した場合、ボード次数は8+8 (入力ポート数+出力ポート数)となり、さらに8ノード実装ではボード次数14+14になる[4]。

RWC-1では、RICAによる細粒度処理を効率良く行なうために、プロセッサでのパケット生成/消費速度と同一のバンド幅が求められており、その値はたとえば48 bits × 50MHz = 2.4 Gbps (300M bytes/sec)となる。相互結合網においてもこれと同等の転送レートが求められるため、必然的に多ビット同時転送となる。基板のエッジから全信号を引き出すと仮定し、有効エッジ長50cm、有効信号ピッチ4本/0.1インチとすると、基板から出せる信号数は約800本と計算される。ボード次数が大きい場合、信号線数がネックとなり、相互結合網のバンド幅を制約することになる。このため、ボード次数を小さく抑えることが必要になる。

そこで、MDCEが実現している良好な転送特性をできるだけ損なうことなく、ボード次数の小さい結合網を得るために、図4に示すようにMDCE網の1ノードを4ノードの小規模なDCE網で置き換えることを考える。図に示すように並行リンクは2重化される。こうしてできた結合網を、本稿では階層型MDCE (c-MDCE)と呼ぶ。たとえばRWC-1 1,024ノードのシステムを構築するためには、$x \times y \times z = 4 \times 4 \times 16$ または $4 \times 8 \times 8$ または $4 \times 16 \times 4$ の構成の(1,1,1)-MDCE網に対して図4に示す置換を行なえばよい。その結果得られる結合網トポロジは、図5に示すようなものとなる。

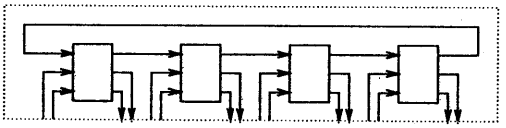

図3: MDCE網のボード次数

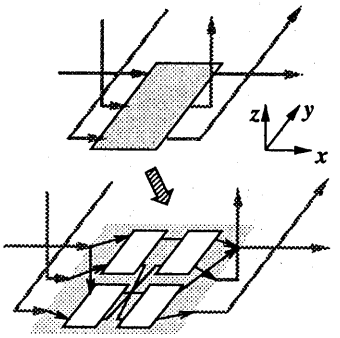

図4: ノードの階層化

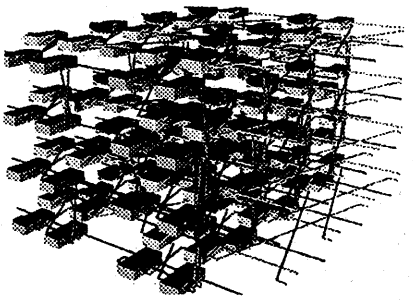

図5: c-MDCE網の概観

[4] 1,024ノードシステムの場合

3.2 ルーティングアルゴリズム

以下に、(1,1,1)-MDCE網を図4のように4ノードで階層化した場合のルーティングアルゴリズムを示す。便宜上、ノードのアドレスは元のMDCE網のx,y,z座標を用い、さらにクラスタ内アドレスwを追加した4次元表現を用いることにする。パケットの送り先アドレスを(w_d, x_d, y_d, z_d), ルータのアドレスを(w_r, x_r, y_r, z_r)としたとき次のようになる。

1. パケットが到着していればプロセッサに渡す
   if $(w_d, x_d, y_d, z_d) = (w_r, x_r, y_r, z_r)$ then
     forward to PE
   endif

2. パケットが目的リングに到着している場合、wアドレスに応じた適切な並行リンクを選択する
   if $y_d = y_r$ and $z_d = z_r$ then
     case (w_r)
     0,2: if $w_d^1 = w_r^1$ forward to parallel-link
          else forward to cross-link
     1,3: forwatd to parallel-link
     endcase
   endif

3. 当該クラスタでc-Banyanホップをする場合、$w=1$のノードに向かう
   if $y_d^{x_r} \neq y_r^{x_r}$ then
     case (w_r)
     0: forward to parallel-link
     1: forward to cross-link
     2: forward to cross-link
     3: forward to parallel-link
     endcase
   endif

4. 当該クラスタでCCCホップをする場合、$w=3$のノードに向かう
   if $z_d^{x_r} \neq z_r^{x_r}$ then
     case (w_r)
     0: forward to cross-link
     1: forward to parallel-link
     2: forward to parallel-link
     3: forward to cross-link
     endcase
   endif

5. それ以外の場合、並行リンクに出力
   forward to parallel-link

ここでx^iはxを2進数表現した時の第iビット(i=0,1,2,...)を表す。

3.3 デッドロック防止

階層化したことにより、もとのMDCE網でCCC型接続をしていた部分で図6にしめすような有向閉路が新たにできる。この部分でstore&forwardデッドロックが起きる可能性があるため、新たにの防止措置を組み込む必要がある。virtual cut-throughルーティングを用いるとき、最小限のチャネル数の増加でデッドロックを回避するためには、たとえば$w=2$においてクロスリンクから入ってきたパケットを(当該ノードが送り先でない限り)必ずクロスリンクから出力する。

上述のようなルーティング制限を行なうことで、新たな有向閉路を禁止した場合、サーキュラ・オメガを含むDCE網やMDCE網で用いられている螺旋バッファ法がそのまま適用できる。すなわち、パケットが$x=0 \rightarrow x=1$のリンクを横切るときにバーチャルチャネル番号を1だけ増す。そのために必要なバーチャルチャネル数は、パケットのリング方向の最大周回数で決まる。c-MDCE網では、階層化によって最大周回数が増すために4チャネル必要になる。

4 評価

4.1 静的特性の評価

表1に階層型MDCE網を含めたいくつかの直接網の次数、ボード次数、bisection width、直径、平均距離を示す。表中の数値はすべて1,024ノードシステムを構成した場合の値である。DCE, MDCE網の直径/平均距離は[7]により解析的に求めたものである。c-MDCE網の直径, 平均距離は上述のデッドロック防止法を加えたルーティングアルゴリズムを用いる場合のものであり、解析が複雑になることからプログラムによって総当り方式で求めている。

表から明らかなように、c-MDCE網は直

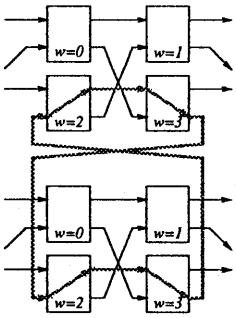

図 6: 階層型 MDCE 網での新たな有向閉路

径、平均距離が大きくなるものの、次数、ボード次数、biection width の各項目とも最小であり、実装性に優れた、低コストのインターコネクションを実現できることがわかる。

4.2 シミュレーションによる評価

実際の転送特性を調べるためにシミュレータを作成し、いくつかの人工的な転送パターンの上での転送性能を測定した。シミュレータではスイッチ、バッファなど主要部分が共通化されており、ルータ間接続方法とルーティングアルゴリズムを記述するだけで、トポロジ等による相違を他の条件を保ったまま比較できる。

シミュレーションは、(1) 1,024 ノードシステム、(2) virtual cut-through ルーティングを行ないチャネル数はデッドロック防止のための最低限のみ用意、(3) 各バーチャルチャネル毎に入力側に 32 ワードのバッファを用意、の条件を統一して行い、ネットワーク上にパケットが無い状態から始めて、10,000 クロック間に目的ノードに到着したパケット数と到着までに要した時間 (クロック数) を測定した。

図 7にランダム通信特性を、図 8に 32×32 メッシュをエミュレートした時の通信特性を、図 9に全体の 5% のパケットが特定ノードに向かう (残り 95% はランダム) 時の通信特性を示す。いずれも横軸がシミュレーショ

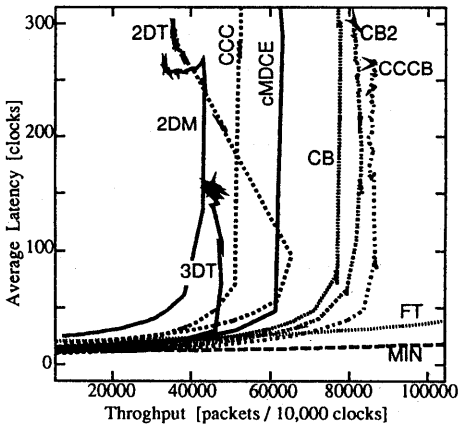

図 7: ランダム通信特性

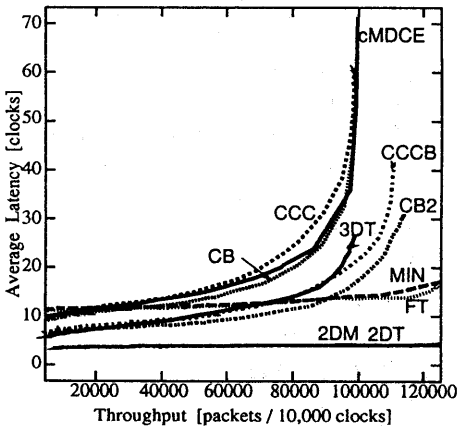

図 8: メッシュエミュレーション通信特性

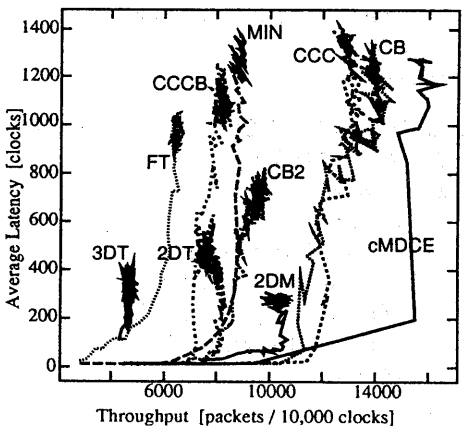

図 9: 5% ホットスポット通信特性

表 1: 各結合網トポロジの静的特性の比較

network	configu-ration	node degree	PCB degree		bisection width	diameter	mean distance
			4 nodes	8 nodes			
c-MDCE	(4)×4×8×8	2+2	4+4	6+6	64	22	12.44
MDCE	4×16×16	3+3	8+8	14+14	256	11	6.44
DCE	8×128	2+2	5+5	8+8	128	14	10.01
2D-torus	32×32	4+4	8+8	12+12	128	32	16.00
3D-torus	8×8×16	6+6	16+16	24+24	256	16	8.00

ン時間 10,000 クロック間で配送されたパケット数（スループット）を示し、縦軸は配送されたパケットの平均配送時間（レイテンシ）を示している。パケット転送量を増すと結合網が輻輳するために平均レイテンシが増す。一定以上のパケット量に対しては網が飽和するためグラフは逆 L 字型になる。図中、CCCB,CB2 は 3 次元 MDCE 網 (1, 1, 1)-, (2, 0, 1)-MDCE 網を、CB, CCC は 2 次元の DCE 網 c-Banyan, CCC 網を、2DM, 2DT, 3DT は 2 次元 /3 次元のメッシュ / トーラス網を示す。FT, MIN は間接網の fat-tree [2], オメガ網 [1] である。なおここでは、実装条件の違いをシミュレーション結果に反映させるため、転送パケット数を 1 ボードに 8 ノード実装した場合のボード次数で正規化している。

図から、DCE, MDCE, c-MDCE の DCE 網族が高いスループットを達成していることがわかる。c-MDCE 網の転送特性は同族の網と比較しても遜色ない性能を出している。グラフ中には直接表現されないが、c-MDCE 網によればリンクあたりのバンド幅を高く保つことが可能であり、この点も含めて c-MDCE 網のメリットは大きいといえる。

4.3 検討

上記評価で用いている 1,024 ノード c-MDCE 網のベース MDCE 網の構成には $x \times y \times z = 4 \times 4 \times 16$, $4 \times 8 \times 8$, $4 \times 16 \times 4$ の 3 通りがあり、構成によって得られる特性が異なってくる。表 2に各構成での最大スループットを示す。表中の最初の 3 項目 (VC=4) はルーティングアルゴリズム、デッドロック防止法に 3.2, 3.3で示した方法を適用した場合である。また、バーチャルチャネルを 8 本用意した場合も (VC=8) として例示する。

図 10は、3.2に示したルーティングアルゴリズムを用いたときのノード間各ポートの使用効率を表している。図では単方向リング 1 つ分だけ抜き出して表示している。ネットワークが飽和状態のとき、シミュレーション時間中、ポート上に有効なパケットが乗っている時間の割合を示している。

図によればノードの位置やポートによってかなりのバラつきがあることがわかる。図は輻輳状態での測定結果であり、低使用度のポートは転送すべきパケットがないため遊んでいるのではなく、受け側ルータにバッファの空きがないためにハンドシェーク機構によって待たされているのである。c-MDCE 網はこの点で改善の余地がある。

表 2: 飽和時のスループット比較

configuration (# of VCs)	throughput
(4)×4×4×16 (VC=4)	68.2K
(4)×4×16×4 (VC=4)	73.1K
(4)×4×8×8 (VC=4)	79.8K
(4)×4×8×8 (VC=8)	82.1K

5 おわりに

本稿では、すでに超並列向けとして提案している直接網 MDCE (Multidimensional Directed Cycles Ensemble extension) で問

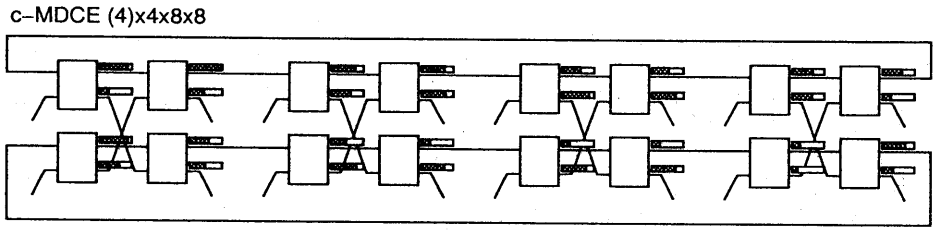

図 10: c-MDCE 網での通信の非均一性

題となる実装性を大幅に改善できる階層型 MDCE 網 (c-MDCE) を提案し、その次数, 直径などを評価するとともに、シミュレータによる転送特性の測定結果を示した。c-MDCE 網は、MDCE 網の各ノードを、4 ノードからなる小規模な DCE 網で置き換えることによって得られる。

c-MDCE 網は次数 2+2 の網でありノード単体での信号線数が少なくてすむ。さらに複数ノードを搭載するプリント基板から引き出される信号数を他の直接網に比べて少なく抑えることができる。したがって c-MDCE 網の採用により超並列計算機の実装性は大幅に向上する。ルータの VLSI ピン数制限、プリント基板から出せる信号数の制限、あるいは筐体間を結ぶワイヤの本数などの物理的制約を満足することができ、その上通信バンド幅を必要なだけ確保することが可能になる。

転送特性は階層化のため元の MDCE 網、DCE 網に比べて若干悪化するが、基本的な特性 / 挙動は元の MDCE 網から継承していると考えられ、メッシュ等の直接網に対する優位性は保たれている。

今後、この c-MDCE 網の採用を前提に超並列計算機 RWC-1 の開発を進めて行く予定である。

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、日頃より有益な議論を頂いている RWC 関係各位に深く感謝します。

参考文献

[1] D. H. Lawrie. Access and Alignment of Data in an Array Processor. *IEEE Trans. Comput.*, Vol. C-24, No. 12, pp. 1145–1155, Dec. 1975.

[2] C. E. Leiserson. Fat-Trees: Universal Networks for Hardware-Efficient Supercomputing. *IEEE Trans. Comput.*, Vol. C-34, No. 10, pp. 892–901, Oct. 1985.

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

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

[5] 坂井修一ほか. 超並列計算機 RWC-1 の基本構想. 並列処理シンポジウム JSPP '93, pp. 87–94, 1993.

[6] 横田隆史ほか. RWC-1 相互結合網用プロトタイプ・ルータの設計. 信学技報, CPSY95-37, June 1995.

[7] 横田隆史ほか. 超並列向け相互結合網 MDCE の提案と評価. 情処学論, Vol. 36, No. 7, pp. 1600–1609, July 1995.