超並列向け相互結合網 MDCE の提案と評価

横田隆史† 松岡浩司† 岡本一晃†
廣野英雄† 坂井修一†

超並列計算機の実現，特に，通信レイテンシの隠蔽や並列性の自然な抽出において優位性を持つマルチスレッド・アーキテクチャでの超並列計算機を考える場合，細粒度・高頻度で非整列なパターンの通信に対する耐性や，通信レイテンシの短縮について相互結合方式を検討しなければならない．また一方で OS などの運用面での検討も必要である．本論文では，まずこのような超並列向けの相互結合網の要件を整理し，次に，間接多段網のスイッチを演算ノードに置き換えて得られるサーキュラ・オメガ網の特質に着目し，その定義を一般化することにより直接網のクラス DCE (Directed Cycles Ensemble) を定義する．そして，任意の DCE 網の直積を考えることで多次元に拡張できることを示し，これによって得られる結合網のクラス MDCE (Multidimensional DCE extension) を提案する．代表的な DCE 網，MDCE 網について直径ならびに平均距離の解析を行い，さらに，シミュレーションにより 5 種類の通信パターンについて動的な転送特性の測定を行った結果，本稿で想定している超並列計算機に適用する場合の MDCE 網の優位性が示される．

MDCE: A New Class of Interconnection Networks for Massively Parallel Computing

TAKASHI YOKOTA,† HIROSHI MATSUOKA,† KAZUAKI OKAMOTO,†
HIDEO HIRONO † and SHUICHI SAKAI †

Multithreaded architecture is an important candidate for massively parallel computers. This architecture has advantages in hiding communication-latency and exploiting parallelism, although, it causes successive fine-grain communication and requires short latency. This paper discusses adequate interconnection networks for that architecture. First we define DCE (Directed Cycles Ensemble) networks as a generalized representation of direct networks derived from multi-stage networks. Then we extend DCE networks into multi-dimensional space and define MDCE (Multidimensional DCE extension) networks. MDCE networks match criteria for massively parallel computers: short diameter and mean distance, scalability, self-routing, uniform traffic, multiple paths, and no interferences among disjoint partitions. Static analysis on diameter and mean distance and simulation results reveal the network's advantages.

1. はじめに

並列処理システムの研究は，処理性能向上の要求を主な原動力として推進されてきた．そして現在では，半導体技術をはじめとする諸技術の発達によって，プロセッサ数 1,000 台以上の超並列計算機について現実性をもって論じられるに至っており，実際にこのような超並列計算機を研究・開発する動きが盛んになっている．その研究の柱のひとつが相互結合方式[1]である．

本稿では，汎用用途の超並列計算機の実現のための有望な候補として挙げられているマルチスレッド・アーキテクチャ[2),3)]をノード・アーキテクチャとして仮定し，これに最も適切な相互結合網を考察する．このアーキテクチャでは，細〜中粒度の処理単位（スレッド）を単位に処理を進めることを基本としており，適正な長さのスレッドが十分な数だけあり，かつシステム中適切に分散されていれば，若干のハードウェア・サポート機能により効率の良い並列処理を実現できる．

ノード・アーキテクチャをこのように設定することで，並列処理システム・アーキテクチャとしての相互結合網にひとつの方向性が出てくる．まず，スレッドが処理の単位になるため，ノード間を行き来するメッセージの単位量がスレッドでの処理量に応じて小さく，また高頻度になる．また一方で，汎用用途の前提

† 技術研究組合 新情報処理開発機構 つくば研究センタ
Tsukuba Research Center, Real World Computing Partnership

から，問題の論理的構造がマシンの物理的構造に一致することは考えにくく，むしろ，通信が系内でランダムに近い形に分散している状況で検討すべきである．

以上の観点に立った場合，超並列向け相互結合網として何が求められるべきなのか．

まず実装およびハードウェア量の問題から，直接網について検討する．間接網はクロスバ網をはじめとして良好な転送特性を示すことが知られているが，結合網機能とプロセッサ機能を（VLSI などの形で）融合しコンパクトに実現することが困難である点，また，スイッチのハードウェア・コストが大きい点の問題があり，超並列用として受け入れ難いためである．

直接網は多くの場合，無向グラフとして表現されるが，現実には，物理的に1本の線路を同時に双方向には使えないため，通信方向を切替えながら使うか，あるいは単方向の線路を2組用意することになる．本稿で前提としている細粒度・高頻度のメッセージ通信の環境では，通信方向の切替えに容認できぬオーバーヘッドが生じるため，本稿では網トポロジを有向グラフで表現する．これに伴い，網次数を入次数・出次数の和として論じることとする．

さらに本稿では超並列に向けたスケーラビリティを問題とする．網としての基本的な特性，すなわち，次数，直径，平均距離，スループット，レイテンシ等において優位性を持ち，さらに，大規模化した場合にもその優位性を保てる結合網方式でなくてはならない．一般的に多くの相互結合網では，システムのサイズが大きくなると直径，平均距離が増すが，そればかりでなくメッセージ間での干渉が増すため，通信の性能は相対的に見れば悪化してゆく．本稿ではこの問題を，網次数に自由度を持たせることで解決を計る．

さらに，転送特性ばかりでなく OS など運用面での配慮も必要である．高価な超並列計算資源の運用を効率的にするため，システムの時分割運用（タイムシェアリング）と空間分割（パーティショニング）をサポートすることである．前者はプロセッサ内のみならず結合網においてもプロセス状態の保存と回復を行う機能を提供するものであり，後者は分割された領域間でパケット転送の干渉を起こさないことを意味する．

以上，本稿で求める超並列向け相互結合網の要件をまとめると次のようになる．

1. 直径が小さいこと，また大規模化に伴う直径の増加も抑えられること．
2. 次数が小さいこと，ただし次数は入次数および出次数の和で数える．
3. 次数に自由度を持ち，規模に応じて増減させることにより，転送特性のスケーラビリティを確保できること．
4. セルフルーティングが可能で，しかも，簡素かつ最適であること．
5. 通信の負荷が均等に分散すること．
6. 2つ以上の経路が選択可能であること．
7. システムが複数の領域に容易に分割可能で，しかも領域間で通信が干渉しないこと．

以降，次章でサーキュラ・オメガ網のクラスが上記要件に良く適合することを示し，それを一般化することにより新しい直接網のクラス DCE（Directed Cycles Ensemble）を定義する．そして第3章で DCE 網を多次元に拡張することによって MDCE（Multidimensional DCE extension）を定義し，第4章で直径および平均距離の解析を行う．また第5章においてシミュレーションによる通信特性の評価を行う．

2. DCE（Directed Cycles Ensemble）網

既存の相互結合網の中で上の要件によく適合するものとして，EM-4 で採用されているサーキュラ・オメガ網[4]に代表される多段型の直接網が挙げられる．この網は，間接多段網（オメガ網[5]）のスイッチをプロセッサ・ノードに置き換え，出力側をそのまま入力にラップアラウンドさせることによって直接網化したものである．この結合網は，直接網であるため網を構成するためのハードウェア負担が少なくてすむほか，網次数・直径ともに小さいなど，前述の要件のほとんどを満たし，本論文で前提としている並列計算機の相互結合網として好ましい特性を持つ．なお，本論文では単に取扱いの簡便さから，Banyan 網[6]を同様の方法により直接網化したサーキュラ Banyan（以降 c-Banyan と略記，**図1**参照）を用いることにする．

ところで文献4)では，サーキュラ・オメガ網を単に直接網化した多段網として扱うのではなく，入力→出力方向に沿って一周する閉路（単方向リング）があることに着目し，興味ある特質を引き出している．c-Banyan 網においても図1から明らかなように横方向に並べられたノードが単方向リングを構成している．これは，見方を変えれば，n 個の要素からなる単方向リング 2^n 組を相互に結合した形と説明することができる．

そこで，c-Banyan 網の構成を次のように定式化することができる．ノードを2次元平面上に $n\times 2^n$ 個並べ，位置 (x, y) で識別するとき，

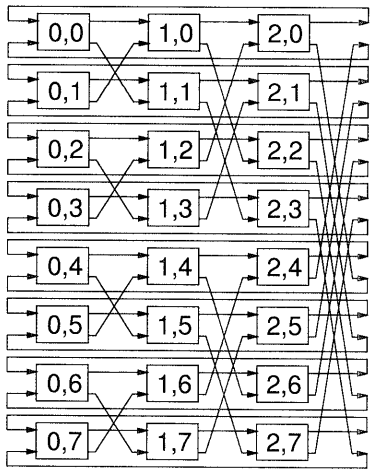
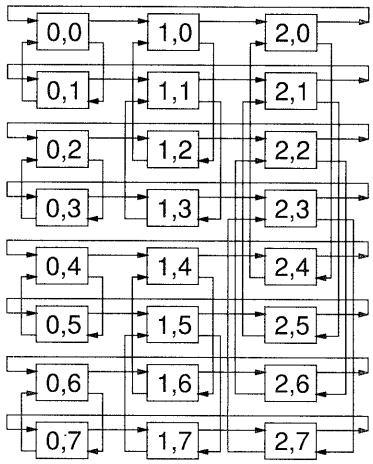
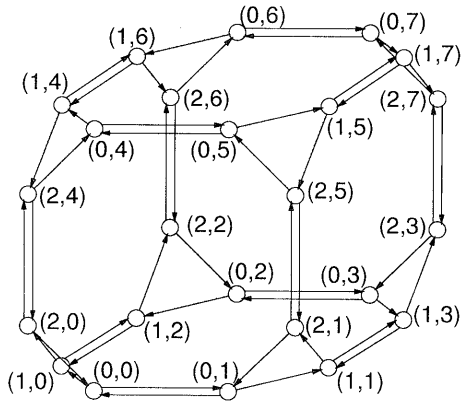


図1 Circular Banyan 網の例
Fig. 1 3×8 Circular Banyan network.



（a） DCE 表現による CCC
（a） CCC expressed by DCE.



（b） （a）と同一の網の別の表現
（b） Another expression of network in (a).

図2 CCC (Cube-Connected Cycles) 網の表現
Fig. 2 CCC (Cube-Connected Cycles) as a DCE network.

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

のように接続する．ここで，$0\leq x<n$, $0\leq y<2^n$ であり，$x \bmod y$ は x を y で割った剰余を表す．平行リンクは x 軸方向に単方向リングを形成し，クロスリンクでリングを相互に接続している．$\delta, f(x, y)$ はクロスリンクを定義するためのパラメータである．$\delta=0, 1, 2, \cdots$ であり，$f(x, y)$ はクロスリンクでの接続関係を表現する関数である．

パラメータの設定によってさまざまな網が定義可能になり，c-Banyan 網を包含する新しい相互結合網のクラスが定義できる．単方向リング（有向閉路）を構成単位として考えることから，この相互結合網クラスを Directed Cycles Ensemble (DCE) と称することにする．また DCE は上の 3 つのパラメータ n, δ, f で表現することができることから，これを DCE(n, δ, f) と表記する．

たとえば $f(x, y)$ を 2 進数表現による y の第 x ビット目の反転 brev とすると，$\delta=1$ のとき網は c-Banyan（図 1）となる．また同条件で $\delta=0$ のとき，網はよく知られた CCC (Cube-Connected Cycles[7]，**図 2**) と同一のトポロジになる（ただしリングが単方向のため，正確には CCC のサブセットである）．

3. MDCE（Multidimensional DCE extension）網

3.1 DCE 網の多次元拡張

DCE 網は文献 4) で示されているように良好な通信特性を持つが，この網クラスは我々が最初に示した要件（3）を満たさない．DCE 網では入次数 2，出次数 2 に固定されているために，網の規模の大きさに対応して直径，平均距離が増加し，それに伴って中継点でのメッセージ・コンフリクトの確率が高くなるために，規模の増加に見合うだけの性能が得られなくなるのである．

そこで，若干の次数の増加を許すことにより，並列規模に見合うだけの転送特性・能力を確保することを考える．次数の増加分は，以下のような方法により DCE 網を多次元方向に拡張するために使う．

まず $n\times 2^n$ DCE 網を 1 枚の平面上に表現する．そして DCE 網が載った平面を垂直方向に 2^n 枚スタックする．このときノードは 3 次元直方体状に $n\times 2^n\times 2^n$ 個並ぶ．ノードの並びを垂直方向に切る x-z 平面には，もとの DCE 網と同様に $n\times 2^n$ 個のノードが並ぶ．

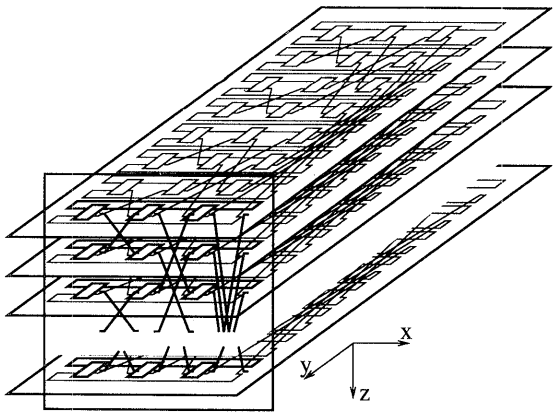


図3 DCE 網の多次元拡張
Fig. 3 Extending DCE network into multidimensional space.

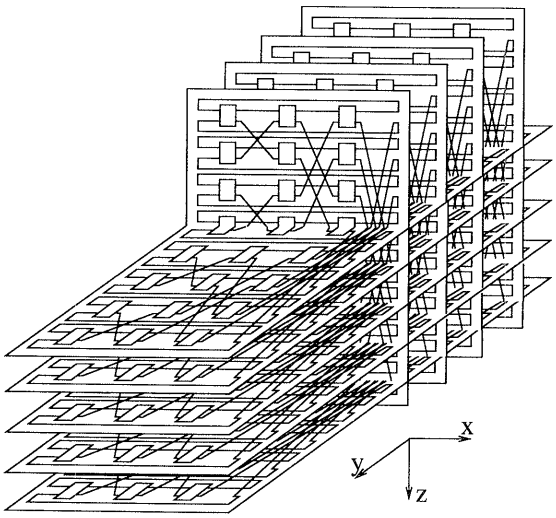


図4 DCE 網の直積
Fig. 4 Cartesian product of DCE networks.

これらのノードも DCE 網で結ぶのである（図3）。

こうして得られる網は x-y 平面で DCE 網を構成し，同時に，x-z 平面でも別個に DCE 網を構成している．平行リンクは x-y，x-z 平面で同一のパターンになるため，物理的には1本に統合してしまうことが可能である．以上の方法によって，入次数1出次数1の増加で全体のトポロジを1次元増すことができる．

このように，網次数を増してノードが配置される空間の次元数を増し，さらに，増した次元数分だけ互いに直交する DCE 網の組で接続する．システム全体での転送特性・能力は，ベースとなる DCE 網の規模と拡張した次元数（＝次数増加分）によって決まる．このように自由度が増すため，許容されるノードの次数や，要求される転送特性などに対して最適な相互結合網を構築することができるようになる．これが DCE 網の多次元拡張の基本的な考えであり，以降，MDCE (Multidimensional DCE extension) と呼ぶことにする．MDCE 網は任意の DCE 網の直積として定義される（図4）．

3.2 代表的な MDCE 網の表現

このように，MDCE 網は任意の DCE 網の直積として定義されるため，高次元の MDCE 網の表現を正確に行おうとすると，もとの DCE 網のパラメータ (n, δ, f) を列挙することになり繁雑である．しかし，マップ関数 $f(x)$ を第2章で用いた brev で代表させ，さらに δ を0または1に制限しても（付録A参照），議論の一般性が大きく損なわれることはない．このため以降では典型的な MDCE 網として元の DCE 網を c-Banyan と CCC の2つに限って検討を進めることにする．

ノードを $(r+1)$ 次元 $(r=2, 3, \cdots)$ に配置した MDCE 網で，平面内のクロスリンクの接続に c-Banyan 接続を用いている次元数を B，CCC 接続を用いている次元数を C，平行リンクの多重度を P とし，これら3つのパラメータで MDCE 網を代表し，(B, C, P)-MDCE と表記する．ここで $B+C=r$ である．

図5は x-y 平面に c-Banyan 網を用い，x-z 平面に CCC を用いた $(1, 1, 1)$-MDCE の概観である．ただしこの図では x 軸方向のラップアラウンド線は省略している．

3.3 ルーティング

(B, C, P)-MDCE において，$r=B+C$ とし，ノード位置を $(r+1)$ 次元の座標 $(x_0, x_1, \cdots, x_r)$ で表す．ルーティング・アルゴリズムは c-Banyan 網からの自然な拡張で，次のようになる．

パケットの送信先アドレスを $(\omega_0, \omega_1, \cdots, \omega_r)$，中継ノードのアドレスを $(x_0, x_1, \cdots, x_r)$ としたとき，各座標についてビットごとの排他的論理和を求め，d とする．$d_i=\omega_i \oplus x_i\ (1 \le i \le r)$ である．

中継ノードは各 d_i の第 x_0 ビットを見て，“1”になっているものを探す．“1”があれば，パケットを対応する出力ポートから出力し，ない場合はそのまま平行リンクから出力すれば良い．

store & forward デッドロックを防ぐための機構も，DCE 網で行われている方法を使うことができる．すでに CCC 網については文献8)で示されており，サーキュラ・オメガ網では文献4)で螺旋バッファ法が提案されている．ただし MDCE 網の場合は，c-Banyan 接続による次元数が多いほど単方向リングに沿っての周回数が多くなるため，これに応じたバッファ構成が必要となる．上記 B, C, P のパラメータを用いれば，

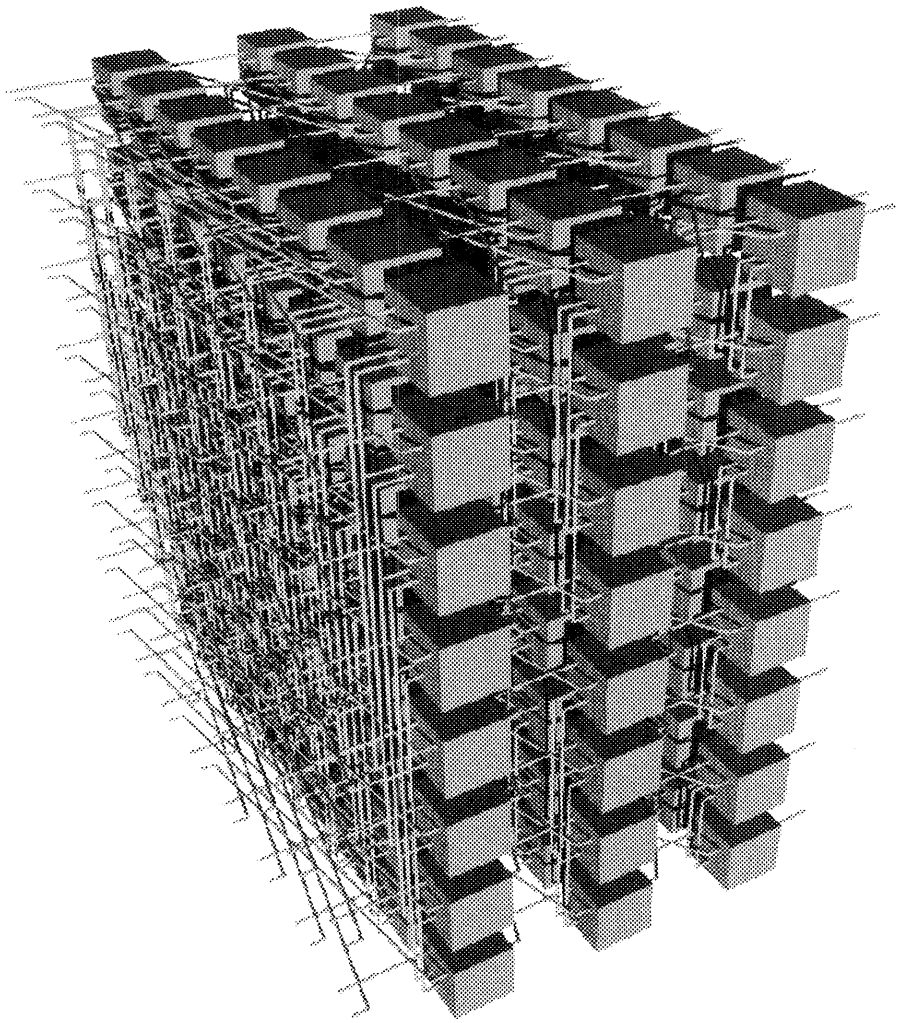


図5 (1, 1, 1)-MDCE 網
Fig. 5 A (1, 1, 1)-MDCE network.

最大の周回数は $(B+1)$ 周であり，そのため螺旋バッファ法を用いる場合は各ノードで $(B+2)$ 個のチャネルを用意する必要がある．

3.4 パーティション

c-Banyan 網は**図6**に示すように y 成分によって 2^m $(m=0, 1, 2, \cdots, n)$ 個の単方向リングを含む領域に分割される．単方向リングを単位としながらこのように分割することにより，領域間での通信の干渉をなくすことができる．

MDCE 網において違う次元に属する DCE 網は互いに直交するため，一方での領域分割の影響を他方が受けることはない．このため，パーティション分割法は DCE 網からの素直な延長となる．すなわち，単方向リングを単位とし，次元 x_i $(i=1, 2, \cdots, r)$ 成分によって

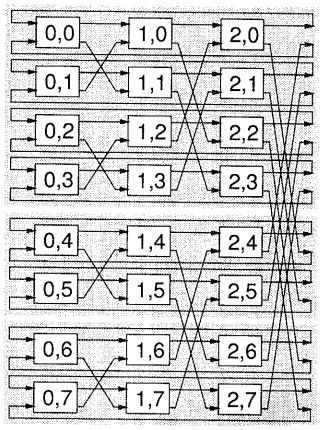


図6 c-Banyan 網の空間分割例
Fig. 6 Partitioning in a c-Banyan network.

2^m 個の領域に分割する．各次元は直交するため，分割軸の選択は x_i の任意の組合せも可能である．

4. 静的特性の解析

4.1 直　　径

c-Banyan 型に接続されたクロスリンクに出力されたパケットは受信側で x_0 位置が変わる．上記ルーティングアルゴリズムで，$d_i, d_j\ (i \neq j)$ の第 x_0 ビットがどちらも“1”である場合，パケットは第 i, j 両ポートのどちらかから出力されることになるが，c-Banyan 型接続では x_0 位置が変わってしまうため，同パケットがリング方向に1周回って元の x_0 位置まで戻ることになる．これに対して CCC 型接続では x_0 位置が変化しないため，パケットが余分な周回をすることはない．ただし，CCC 型のクロスリンクのみを持つ MDCE 網では，x_0 位置を変えるために必ず平行リンクを通る．

そこで，直径は B の値により次のように求められる．

$$\text{直径}=\begin{cases}(B+C+1)n-1 & (B>0)\\ (C+2)n-2 & (B=0)\end{cases} \qquad (1)$$

なおこの式は $B=0$（または $C=0$）の DCE 網にも適用できる．

4.2 平均距離

まず，DCE のクラスに属する c-Banyan 網について考える．3.3 節で示したルーティング方法をとるとき，パケット送出ノードの y アドレス y_s と行き先の y アドレス y_d のビットごとの排他的論理和 b が以降の配送経路を表現する．まず送出ノードで b の第 x_s ビットを見て“1”ならばクロスリンクに，“0”ならば平行リンクに転送する．これを受けたノードでもまた同様のことが行われ，パケットが転送される．パケットは単方向リングに沿って1周以内に目的のリングに到達し，平行リンクのみを伝わって目的のノードに配送される．

平均距離を考える場合，対称性から最初に見るビット位置 (x_s) を考慮する必要がない．そこで，$b=y_s \oplus y_d$ のなかで“1”となっている最上位のビットの位置を $j\ (0 \leq j < n)$ とする．第 j ビットが“1”になっていることから，少なくとも j ホップしないと目的のリングに到達しないことがわかる．目的リングに着いてから送り先ノードに着くまで平均リング半周分のホップ数を要するから，平均距離は結局，

平均距離（c-Banyan）

$$=\frac{1}{2^n}\sum_{j=0}^{n-1}((j+1)2^j)+\frac{n-1}{2}$$

$$=\frac{3}{2}(n-1)+\frac{1}{2^n} \qquad (2)$$

となる．

CCC 網についても 3.3 節のルーティングに従う場合の平均距離を求めてみる．上と同様に $b=y_s \oplus y_d$ のなかで“1”となっている最上位のビットの位置を $j\ (0 \leq j < n)$ とする．第 j ビットに相当する位置まで少なくとも j ホップが必要である点は上と同様である．CCC の場合，クロスリンクが x アドレスを変えないため，パケットが第 j ビットに相当する位置に進むまでに平均 $(j/2)$ 回クロスリンクを通過する．このため平均距離は次のようになる．

平均距離（CCC）

$$=\frac{1}{2^n}\sum_{j=0}^{n-1}\left(\left(\frac{3}{2}j+1\right)2^j\right)+\frac{n-1}{2}$$

$$=2n-\frac{5}{2}+\frac{1}{2^{n-1}} \qquad (3)$$

MDCE 網についても，上記 DCE 網と同様の方法により平均距離が求められる．

まず各座標成分ごとに排他的論理和を求める．$\boldsymbol{v}$ を送出ノードのアドレス，$\boldsymbol{w}$ を送り先アドレスとしたとき，各座標成分ごとにビットごとの排他的論理和 $d_i=w_i \oplus v_i\ (1 \leq i \leq r)$ を求める．そして $d_i\ (1 \leq i \leq r)$ 内にある“1”の数を，ビット位置ごとに集計し

$$c_j=\sum_{i=1}^{r}(d_i)_j \cdot \delta_i \quad (0 \leq j < n) \qquad (4)$$

を求める．ここで δ_i はクロスリンクの接続パターンを示し，c-Banyan 型の場合1，CCC 型の場合0とする．$(d_i)_j$ は第 i 次元座標の第 j ビットを表す．

上記 c_j の中の最大の値を k とすると，パケットはリング方向に i ホップ進み，そこからリングに沿って $k-1$ 周回し（$(k-1)n$ ホップ），さらに平行リングを平均半周（$(n-1)/2$ ホップ）して目的地に到着する．さらに CCC 型接続によるホップ（平均 $nC/2$）も加わる．全体で 2^{nB} だけの組合せがあり，そのなかで上記の場合の数は

$$g(n, B, k, i)$$

$$=\left(\sum_{j=0}^{k}\binom{B}{j}\right)^{i-1}\times\binom{B}{k}\times\left(\sum_{j=0}^{k-1}\binom{B}{j}\right)^{n-i} \qquad (5)$$

である．以上から平均距離を算出することができ，

平均距離（MDCE）

$$=\frac{1}{2^{nB}}\sum_{k=1}^{B}\sum_{i=1}^{n}\left\{\left((k-1)n+i+\frac{nC}{2}\right)\times g(n, B, k, i)\right\}+\frac{n-1}{2} \qquad (6)$$

となる．

表1 は，上で求めた平均距離を 1,024 PE，512 KPE

表 1 網の次数ならびに平均距離の比較
Table 1 Comparison on degree and mean distance.

Network topology	degree (IN+OUT)	1,024 PEs		512 K PEs	
		config./mean dist.		config./mean dist.	
(2,0,1)-MDCE ((CB)²)	3+3	4×16×16	8.15	8×256×256	18.61
(1,1,1)-MDCE (CCCB)	3+3	4×16×16	6.44	8×256×256	14.49
c-Banyan (CB)	2+2	8×128	10.01	16×32768	22.00
Cube-Connected Cycles (CCC)	2+2	8×128	12.52	16×32768	28.50
2D Mesh (2DM)	4+4	32×32	32.00	1024×512	768.00
2D Torus (2DT)	4+4	32×32	16.00	1024×512	384.00
3D Torus (3DT)	6+6	8×8×16	8.00	64×64×128	64.00
binary Fat-Trees (FT)	4+4	1024×10	17.00	2^{19}×19	35.00
Omega (MIN)	2+2	512×10	10.00	2^{18}×19	19.00

システムについて算出し，他の結合網と比較したものである．CB，CCC は DCE 定義による c-Banyan，Cube-Connected Cycles を指す．また CCCB (Cube-Connected Circular Banyans) は (1, 1, 1)-MDCE の別名であり，$(CB)^2$ は (2, 0, 1)-MDCE の別名である[9]．

5. シミュレータによる評価

転送特性の評価のためのシミュレータを作成し，評価を行った．シミュレータは C++言語で書かれており，スイッチ，バッファなどは共通部品化されている．このため網次数など若干のパラメータを設定し，ルータ間の接続のしかた（トポロジ）と，ルーティングアルゴリズムを記述すれば動作する．トポロジ等による違いを，他の条件を全く同一に保ったまま比較できるようになっている．

シミュレーションは以下に示す仮定を統一して行った．

（a）システムの PE 数は 1,024．

（b）実装上，一度に転送される量が LSI のピン数により制約されるものとした．リンクあたりの転送幅が次数によって決まることから，1 パケットの転送に（入次数＋出次数）の仮想クロックを要するものとした．

（c）virtual cut-through で決定的 (deterministic) ルーティングを行う．チャネル数はトポロジにより store & forward デッドロックを起こさない最低限だけに限る．

（d）バッファはチャネルあたり 32 ワードで固定．

また，シミュレーション評価は，以下の 5 つの通信パターンに関して行った．

（1）ランダム通信
全くランダムな相手に対してパケットを送る．

（2）1/4 パーティション通信
全体を同じ大きさのパーティション 4 個に切り，パーティション内でランダムに通信する．

（3）ホットスポット通信
ランダム通信と同様．ただし，全送出パケットの 5%が特定の PE 宛になる．

（4）メッシュ・エミュレーション
PE をアドレスをもとに単純に 32×32 のメッシュにマップし，隣接 4 PE 宛に 1 個ずつパケットを送り，4 個のパケットを受信するまで待つ．

（5）近接通信
ランダム通信と同様だが，通信相手を選ぶ際，アドレス距離が近いほど通信する確率を高くする．通信相手との座標成分ごとの差が当該座標成分の最大値の 1/2 を平均値とする指数分布となるように選んだ．

これらの結果を**図 7～11** に示す．図中，CB は c-Banyan を，2 MD，2 DT，3 DT は各々 2 次元メッシュ，2 次元トーラス，3 次元トーラスを，FT は完全バイナリ fat-trees[10] を，MIN は（間接の）オメガ網をそれぞれ表す．$(CB)^2$ は (2, 0, 1)-MDCE の別名であり，CCCB は (1, 1, 1)-MDCE の別名である[9]．なお，DCE 網 (c-Banyan，CCC) ではノード数を他と合わせるため，完全網構成 7×128 に 1 段追加した 8×128 の構成を用いている．これは他が完全網構成を取るため評価上やや不利に働く．

各グラフにおいて，x 軸はシミュレーション時間 (10,000 仮想クロック) 内に配送されたパケットの総数を示し，log スケールになっている．y 軸はパケットの平均レイテンシである．

平均レイテンシは，網が込んできてスループットが高くなると次第に大きくなり，ある閾値を越えると一気に上昇する．このためグラフは一般に逆 L 字型になる．この閾値が網のトラフィック耐性を示す．むろん高いスループットまで耐えられる網が望ましい．

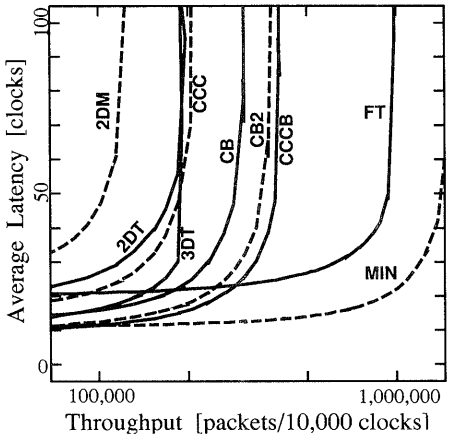


図7 ランダム通信特性
Fig. 7 Random traffic performance.

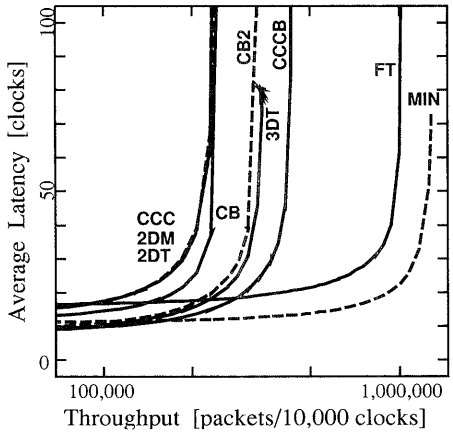


図8 1/4パーティション通信特性
Fig. 8 1/4 Partition performance.

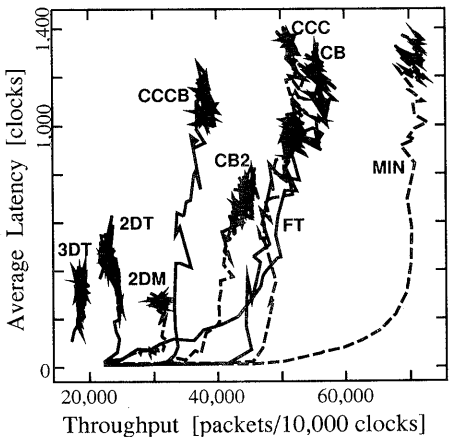


図9 5%ホットスポット通信特性
Fig. 9 5% Hot-Spot traffic performance.

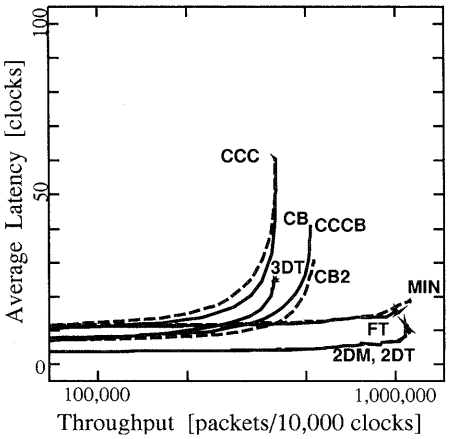


図10 メッシュエミュレーション特性
Fig. 10 Mesh-Emulation performance.

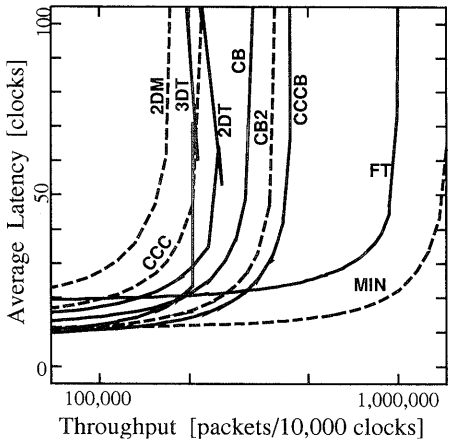


図11 近接通信特性
Fig. 11 Localized traffic performance.

各グラフ中，間接網（図中 FT，MIN）は，いずれの通信パターンについても高いトラフィック耐性を示している．しかしここでの評価にはハードウェア量が勘案されていないため，直接に比較の対象とはならない．

図7，8，11から，MDCE のクラスの網（$(CB)^2$ および CCCB）の特性は，トラフィック耐性において間接網に及ばないが，非飽和状態でのレイテンシは最小レベルであり，次数，ハードウェア量などとのトレードオフで考えれば MDCE 網の他の網に対する優位性が理解されよう．

図9は各網のホットスポット特性を示している．ホットスポットが存在する転送パターンの下では，輻輳が局所的にとどまらず広い範囲に拡大してしまう tree saturation[11] が起きる．試行ごとのばらつきが大きいため，グラフではプロットをあえて線で結んでい

る．同図から(M)DCE網や間接網ではメッシュなどに比べ平均レイテンシが著しく大きくなることがわかる．これらの網は，直径がおおよそ $O(\log N)$ であるなど良好な転送特性を持つが，このことが逆にtree saturationの拡大を助け，輻輳がほぼ一様に拡がってしまうためと考えられる．逆にメッシュ等では，tree saturationの影響を受けにくい部分が残るために，この輻輳のひどくない部分で通信が行われる．このために平均のレイテンシは短くなる．

MDCE網はトポロジ上，メッシュを直接マップすることはできないが，メッシュ・エミュレーション時においても，メッシュ形状に次いで良好な特性を示している(図10)．ここでトポロジ・マッピングはノードアドレスから単純に機械的に行ったものであり，トポロジごとに考えられる最適化は考慮していない．一方でDCE網ではトポロジのミスマッチによる性能低下が端的に現れていることが読み取れる．

6. おわりに

本論文では，サーキュラ・オメガのクラスの直接網の定義を一般化し，複数の単方向リングから構成される相互結合網のクラスDCE (Directed Cycles Ensemble)を提案し，さらに，DCE網を多次元拡張した結合網のクラスMDCE (Multidimensional DCE extension)を提案した．MDCE網は任意のDCE網の直積によって得られる．MDCE網は，直径，平均距離，次数といった相互結合網としての基本的特性をDCE網から継承している．さらにこの網は次数に自由度が与えられるために，十分なスケーラビリティを保ったまま大規模化に対応することができ，超並列向きと言える．

DCE網，MDCE網の代表的なケースについて直径ならびに平均距離を解析し，併せて，シミュレータにより動的な転送特性も評価した．これにより，MDCE網が超並列向け相互結合網として有望であることが示された．

今後，さまざまな通信パターン，特に実アプリケーションでの挙動・特性を評価するとともに，実装方法やOS支援などの付加機能などについても検討を進め，実マシンRWC-1[12]への実装を行い，有効性を検証する予定である．

謝辞 本研究の機会を与えて頂いたRWCつくば研究センター島田潤一所長，MDCE網の原形となるヒントを頂いた慶應義塾大学の天野英晴助教授，また，有益な議論を頂いたRWC関係各位に深く感謝いたします．

参考文献

1) Wu, C.-L. and Feng, T.-Y. eds.: *Tutorial: Interconnection Networks for Parallel and Distributed Processing*, IEEE Computer Society (1984).
2) Culler, D. E. et al.: Fine-grain Parallelism with Minimal Hardware Support: A Compiler-Controlled Threaded Abstract Machine, *Proc. 4th Intl. Conf. on Architectural Support for Programming Languages and Operating Systems (ASPLOS-IV)*, pp. 164-175 (1991).
3) Sakai, S. et al.: Reduced Interprocessor-Communication Architecture for Supporting Programming Models, *Proc. Conf. on Massively Parallel Programming Models*, pp. 134-143 (1993).
4) Sakai, S., Kodama, Y. and Yamaguchi, Y.: Design and Implementation of a Circular Omega Network in the EM-4, *Parallel Computing*, Vol. 19, No. 2, pp. 125-142 (1993).
5) Lawrie, D. H.: Access and Alignment of Data in an Array Processor, *IEEE Trans. Comput.*, Vol. C-24, No. 12, pp. 1145-1155 (1975).
6) Goke, L. R. and Lipovski, G. J.: Banyan Network for Partitioning Multiprocessor Systems, *Proc. 1st Ann. Symp. on Computer Architecture*, pp. 21-28 (1973).
7) Preparata, F. P. and Vuillemin, J.: The Cube-Connected Cycles: A Versatile Network for Parallel Computation, *Comm. ACM*, Vol. 24, pp. 300-309 (1981).
8) Dally, W. J. and Seitz, C. L.: Deadlock-Free Message Routing in Multiprocessor Interconnection Networks, *IEEE Trans. Comput.*, Vol. C-36, No. 5, pp. 547-553 (1987).
9) 横田ほか：超並列計算機RWC-1の相互結合網，情報処理学会アーキテクチャ研究会，ARC-101-4 (1993).
10) Leiserson, C. E.: Fat-Trees: Universal Networks for Hardware-Efficient Supercomputing, *IEEE Trans. Comput.*, Vol. C-34, No. 10, pp. 892-901 (1985).
11) Pfister, G. F. and Norton, V. A.: "Hot Spot" Contention and Combining in Multistage Interconnection Networks, *IEEE Trans. Comput.*, Vol. C-34, No. 10, pp. 943-948 (1985).
12) 坂井ほか：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).

付録

A. DCE網の特性について

本文中，DCE(n, δ, f)のうち$\delta=0, 1$の場合のみを

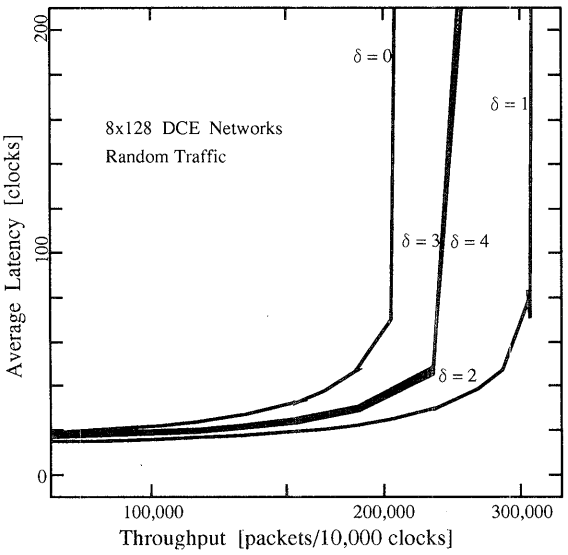


図12 DCE 網転送特性の δ による相違
Fig. 12 Transfer performances for various δ in DCE networks.

扱った点について補足する.

DCE 網は有向グラフで定義され,ノード間のリンクは 2 次元配置の桁位置(x 軸成分)が同じか増える方向に接続されている.パケットの転送経路を考えるとき,$\delta=0,1$ の場合は桁位置が不連続にスキップすることはないが,$\delta>1$ のときはクロスリンクを通ったときに桁位置がスキップする.このスキップにより転送ホップ数が減る場合もあるが,スキップされた桁位置からのクロスリンクを使うときは必ずリングに沿って 1 周してその桁位置に戻らねばならないため,全体で平均すると網直径・平均距離ともに悪化する.このことは**図12**に示すように転送特性の悪化としてシミュレーションによっても検証されている.

また δ が大きくなるとリングに沿っての周回数が増すため,store & forward デッドロック防止のためのチャネル数を多くとらねばならない.これはハードウェア上大きな負担になる.

なお,$\delta=0$ の場合は他に比べ十分な特性が得られていないが,(1)良く知られた CCC 網を表現している点,(2)直積により MDCE 網を構成する際,$\delta=1$ だけの組合せ ((2, 0, 1)-MDCE) よりも $\delta=0,1$ の組合せ ((1, 1, 1)-MDCE) のほうが良好な転送特性を示す点,の理由により評価対象として残している.

(平成 6 年 9 月 16 日受付)
(平成 7 年 3 月 13 日採録)

横田 隆史(正会員)

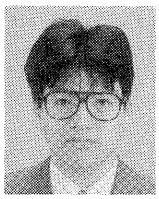
1960 年生.1983 年慶應義塾大学工学部電気工学科卒業.1985 年同大学院電気工学専攻修士課程修了.同年三菱電機(株)入社,知識処理向けアーキテクチャおよび並列アーキテクチャの研究に従事.1993 年 12 月より(技組)新情報処理開発機構に出向.超並列アーキテクチャの研究に従事.電子情報通信学会会員.

松岡 浩司(正会員)

1961 年生.1984 年東京工業大学工学部電気電子工学科卒業.1986 年同大学院理工学研究科修士課程修了.同年,日本電気(株)入社.1992 年 10 月(技組)新情報処理開発機構に出向,研究員.現在,計算機システム一般,特にプロセッサアーキテクチャの研究に従事.

岡本 一晃(正会員)

1962 年生.1986 年慶應義塾大学理工学部電気工学科卒業.同年三洋電機(株)に入社.1992 年 10 月より(技組)新情報処理開発機構に出向.並列計算機アーキテクチャの研究に従事.



廣野 英雄

1968 年生.1991 年筑波大学第三学群基礎工学類卒業.同年三洋電機(株)に入社.1992 年 10 月より(技組)新情報処理開発機構に出向.並列計算機アーキテクチャの研究に従事.電子情報通信学会会員.

坂井 修一(正会員)

昭和 33 年生.昭和 56 年東京大学理学部情報科学科卒業.昭和 61 年同大学院情報工学専門課程修了.工学博士.同年,電子技術総合研究所入所.平成 3 年 4 月より 1 年間米国 MIT 招聘研究員.平成 5 年 3 月より RWC 超並列アーキテクチャ研究室室長,現在に至る.計算機システム一般,特にアーキテクチャ,並列処理,スケジューリング問題などの研究に従事.情報処理学会研究賞(平成元年),同論文賞(平成 2 年度),元岡記念賞(平成 3 年),日本 IBM 科学賞(平成 3 年)各受賞.