

# 超並列向き階層型相互結合網 de Bruijn Connected Torus (BCT) の理論的性能

内角 哲人 堀口 進

uchikado@jaist.ac.jp hori@jaist.ac.jp

北陸先端科学技術大学院大学 情報科学研究科

## 概要

近年、並列計算機のプロセッシングエレメント (PE) を結合するネットワークとして階層型相互結合網が注目されている。また、1枚のウェハ上に多数の PE を実装して、複数のウェハを縦に積み重ねた 3次元スタック構造が実装されるようになってきた。しかし、従来の相互結合網では PE 数が多くなるとネットワークの通信性能が低下し、リンク次数やリンク総数の増加によりネットワークを実装できないなどの問題がある。本稿では、単純な格子網を基本構成 (BM) としてウェハ上に実装し、BM 間をダブルジーン網で結合した階層型相互結合網 de Bruijn Connected Torus を提案する。そして BCT のネットワークの理論性能やレイアウト面積について導出し従来のネットワークと比較し、有効性を示す。

## Network Performance of Hierarchical Interconnection Network: de Bruijn Connected Torus (BCT)

Akihito Uchikado and Susumu Horiguchi

Graduate School of Information Science, Japan Advanced Institute of  
Science and Technology

## abstract

Hierarchical interconnection networks have been attractive networks to construct a large number of processing elements for massively parallel computers in 3D Stacked implementation. However conventional networks have difficult problems to implement a large number of links in 3D stacked. This paper addresses a hierarchical interconnection networks: de Bruijn Connected Torus (BCT). The BCT is a hierarchical network of the de Bruijn whose node is 2D Torus named as a basic module (BM). Network performances of the BCT are obtained theoretically and are compared with conventional networks.

## 1 はじめに

大規模な物理シミュレーション、VLSI CAD や画像処理など科学技術の最先端分野においては現在のスーパーコンピュータより高速なコンピュータが必要とされている。処理能力向上の手段として、多数のプロセッシングエレメント (Processing Element: PE) を相互結合させた超並列計算機が注目され、高速処理に関する研究が盛んに行われている。

超並列計算機システムを実現することは、膨大なレ

イアウト面積、信頼性の低下、コストの増大などの問題が生じるため大変困難である。そこで、超並列計算機システムの一部を一つのチップのように一枚のウェハ上に実装する Wafer Scale Integration (WSI)[1] が注目されている。しかし、ウェハの大きさは技術的に制約されるため、一枚のウェハに実装できる PE 数に制限がある。J.Carson[3] は、ウェハを使って超並列計算機システムを構築する方法として、ウェハを縦に積み重ねたウェハスタック構造を提案している。

ウェハスタック構造はウェハ間の結線数が問題となってくる。従来の相互結合網は、PE数を増やすとネットワーク性能が低下し、リンク次数やリンク総数の増加によりネットワークを実装できないなどの問題がある。Y.R.Potlapalli[2]は、PE数が増加しても相互結合網の直径の増加を抑え、少ないリンク次数を保ち、局所通信を利用できる階層型相互結合網が超並列計算機に適していることを指摘している。

本稿では、実装の容易な2次元Torusを基本構成(Basic Module: BM)とし、直径、リンク次数、拡張性に優れたde Bruijn[4]でBM間を結合する超並列向き階層型相互結合網de Bruijn Connected Torus (BCT)の提案を行う。またBCTのネットワーク性能を理論的に解析し、直径、リンク次数、Bisection Widthやレイアウト面積などを従来のネットワークと比較検討する。

本稿の構成は以下の通りである。第2章では、BCTのネットワーク構成について述べる。第3章ではBCTネットワークの理論性能を導出し、定式化する。第4章ではレイアウト面積について検討する。第5章では、BCTと従来のネットワークと比較検討し、BCTの有効性を示す。第6章はまとめである。

## 2 BCTのネットワーク構成

### 2.1 BCTネットワーク

BCTはレベル2階層( $L_2$ )から最大階層( $L_{max}$ )を構成できる階層型相互結合網である。以下、階層をレベルと呼ぶ。図1に階層型相互結合網BCTの基本的な結合図を示す。 $L_1$ は2DTorusをBMとして用い、 $L_2 \sim L_{max}$ にはde Bruijnを用いてBM間を結合する。各BMは、図1に示すように周りに位置するPEに階層間のリンクを持ち、これらのリンクを用いてBM間を結合する。2DTorusは、実装面を考慮してリンクを双方向だけでなく単方向の2通りを考える。

ここで、2DTorusをde Bruijnでつなげたネットワークを $BCT(n, m, L, q)$ と定義する。 $n$ はde Bruijnの次元数、 $m$ はBMの各辺の次元数、 $L$ は階層のレベル数、 $q$ は結合度を表す。de Bruijn用のリンクは $q$ の値によりどの階層と結合するかが決まる。図2に $m=2, L=L_{max}$ の場合の結合度 $q$ の違いによる階層レベル間結合図を示す。図3に $m=2, q=0$ の場合、 $L=2, 3, 4, 5$ の階層レベル間結合の例を示す。

### 2.2 de Bruijn網

階層型ネットワークBCTの上位階層であるde Bruijn網[4]を $dB(r, n)$ と定義する。ここで、 $dB(r, n)$

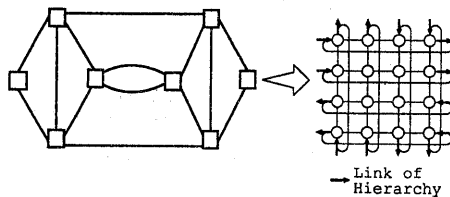


図1: 階層型相互結合網BCTの結合図

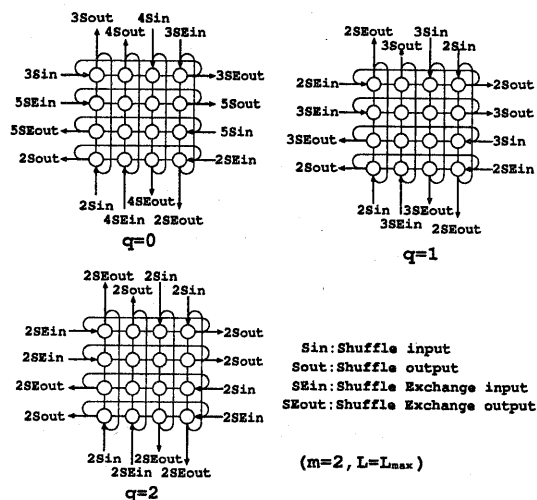


図2: 結合度 $q$ による各BMの階層間リンク結合

の全ノード数は $N = r^n$ である。任意のノードアドレスを $n$ 個の基数 $r$ を用いて $(r_{n-1} \dots r_0)$ と表すと、結合されるノードアドレスは以下のように示される。

$$\text{Input node} \quad (r_{n-2} r_{n-3} \dots r_0 \alpha)$$

$$\text{Output node} \quad (\alpha r_{n-1} \dots r_2 r_1)$$

$$(\alpha = 0, 1, \dots, r-1)$$

$r=2, 3, 4$ の時、de Bruijn網は最も良いネットワーク性能が得られる。本論文では、アドレスの割当の容易さから以下では $r=2$ とする。

## 3 BCTネットワーク性能

### 3.1 理論性能の導出

ここでは、BCTのノード数、直径、リンク次数、コスト、リンク総数、Bisection Widthについて理論性能を導出する。 $BCT(n, m, L, q)$ の全ノード数 $N_{BCT}$ は次式で求められる。

$$N_{BCT} = 2^{2m+n(L-1)} \cdot (2 \leq L \leq L_{max} = 2^{m-q} + 1) \quad (1)$$

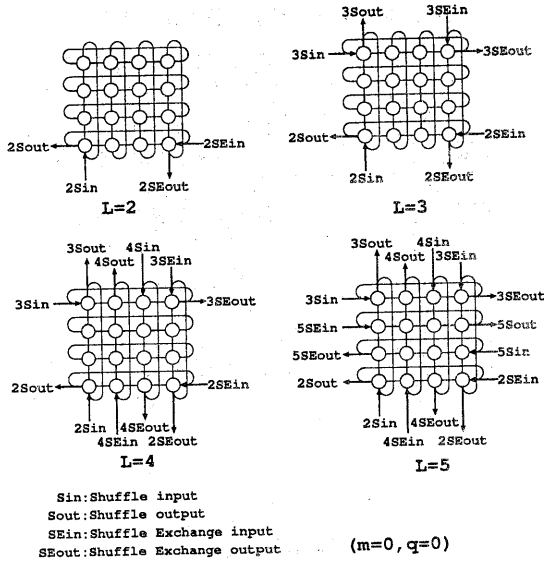


図 3: 各レベル  $L = 2, 3, 4, 5$  の階層間リンク結合

次に BM の送信元の直径を  $DIA_{BMS}$ , BM の送信先の直径を  $DIA_{BMD}$  として各レベルの直径を  $DIA_{TORI(i)}$  とする. BCT の直径 ( $d_{BCT}$ ) は次式に示す総和で表される.

$$d_{BCT} = DIA_{BMS} + DIA_{BMD} + \sum_{i=L}^2 DIA_{TORI(i)} \quad (2)$$

リンク次数  $k_{BCT}$  は, 下位階層ネットワークのリンク次数に 1 を加算したものである. 2DTorus はネットワークサイズによらず 4 と一定であるので, BCT のリンク次数は  $k_{BCT} = 5$  である. これは, ネットワークサイズによりリンク次数が変化しないので, 大規模向きである特徴を示す. 上記からコストは, 双方向リンクに注意して以下のように求められる.

$$COST_{BCT} = \begin{cases} 10DIA_{BCT} & (\text{単方向}) \\ 6DIA_{BCT} & (\text{双方向}) \end{cases} \quad (3)$$

BCT の Bisection Width は, レベル  $L_{max}$  の de Bruijn の Bisection Width の定数倍に等しい. まず 2 進 dB の Bisection Width を以下に示す.

$$BW_{dB} = \begin{cases} (n+1) \frac{\binom{n}{\frac{n}{2}}}{n} & (n = 1, 3, \dots) \\ 2 \left\{ \left( \frac{n-2}{2} - 1 \right) + \left[ \frac{\binom{n}{\frac{n}{2}}}{n} \right] \right\} & (n = 2, 4, \dots) \end{cases} \quad (4)$$

表 1: BCT の Bisection Width:  $n = 4096$

L	$q = 0$	$q = 1$	$q = 2$
2	58	116	232
3	128	256	—

表 2: 各パラメータによる BCT の直径 ( $m=2$ )

L	BCT単方向			BCT双方向		
	$q = 0$	$q = 1$	$q = 2$	$q = 0$	$q = 1$	$q = 2$
1	6	6	6	4	4	4
2	$4n + 9$	$4n + 8$	$4n + 6$	$2n + 7$	$2n + 6$	$2n + 5$
3	$8n + 10$	$8n + 9$	—	$4n + 8$	$4n + 6$	—
4	$12n + 13$	—	—	$6n + 10$	—	—
5	$16n + 16$	—	—	$8n + 13$	—	—

4 式から BCT の Bisection Width は, パラメータ  $q$  と  $L$  によって以下のように与えられる.

$$BW_{BCT} = \begin{cases} 2^q BW_{dB} & L = 2 \\ 2^{q+n} BW_{dB} & L > 2 \end{cases} \quad (5)$$

表 1 に BCT の各パラメータによるノード数 4096 の場合の Bisection Width を示す.

### 3.2 BCT の最適ネットワーク構成

表 2 に BCT の各パラメータ ( $L, q, n$ ) による直径を示す. 表 2 を基に図 4 にレベル  $L = 2$  の BCT の  $q$  をパラメータとしたノード数と直径の関係を示す. 図 4 より  $q = 2$  の場合が直径が最も短い. これは BM で階層間のリンクのポートが  $q = 0$  のときの 4 倍になるためである. 次に図 5 に結合度  $q = 0$  の場合の  $L$  をパラメータとしたノード数と直径の関係を示す. 図 5 より  $L = 2$  のときに最も直径が短い. これは経路する BM 内の通信ステップ数が  $L = 2$  のとき最も少ないためである. Bisection Width は, レイアウト面積のおおまかな見積もりに使われる.  $L = 2, q = 0$  の時,  $BW_{BCT}$  は最も小さく, レイアウト面積も小さくなると予想される.

## 4 BCT のレイアウト面積

### 4.1 レイアウト面積の評価

レイアウト面積は, PE 幅, スイッチ (SW) 幅, ウェハ上の PE を結合するための結線幅, ウェハ間の結線幅から成る. これらのパラメータの一覧を表 3 に示す. 図 6 にウェハ上の PE と SW の配置図を示す. PE は正方形に等しい間隔で並べられると仮定する. SW は正方形に並べられた PE の上下左右に 1 個ずつ配置する. ま

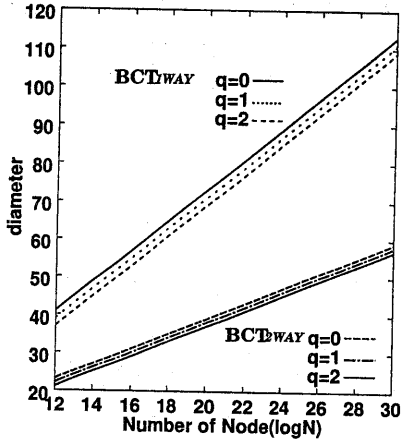


図 4: BCT のノード数と直径の関係 ( $L=2$ )

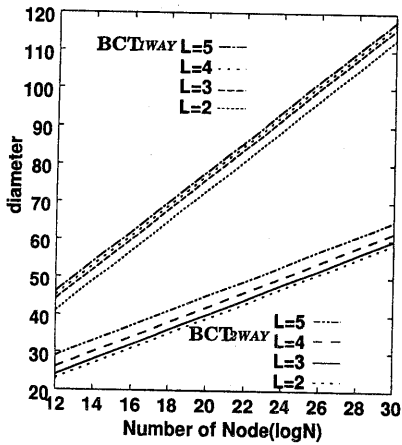


図 5: BCT のノード数と直径の関係 ( $q=0$ )

た SW はクロスバスイッチを用いる。ゴールドバンプ (ウェハ間結合のためのリンク)[5] はウェハの最も外郭に一列に並べられているものと仮定する。ウェハ上に分割した BCT をレイアウトした場合の面積幅  $W$  は次式で与えられる。

$$W = \sqrt{m}(W_{PE} + W_{PElink}) + 2(W_{SW} + W_{SWlink} + W_{GB}). \quad (6)$$

近年の VLSI 技術の進歩により  $1\mu\text{m}$  CMOS technology から  $0.25\mu\text{m}$  CMOS technology へ短くなってきている。また PE も 32 ビットから 64 ビットになったことから PE の面積が大きくなってきている。すなわち、 $W_{PE} \gg W_{PElink}$ 、 $W_{SW} \gg W_{SWlink}$  であること

表 3: レイアウト面積に関するパラメータ

$W$	ウェハの一辺の幅
$m$	ウェハ上の PE 数
$W_{PE}$	PE の一辺の幅
$W_{PElink}$	ウェハ内 PE 結線に必要な結線幅
$W_{SW}$	SW の一辺の幅
$W_{SWlink}$	ウェハ間結合に必要な結線幅
$W_{GB}$	ゴールドバンプの幅
$A_{PE}$	PE の面積
$A_{SW}$	SW の面積
$A_{GB}$	ゴールドバンプの面積
$C_{peak}$	ウェハ間最大結線数
$N_{SW}$	SW の入出力数

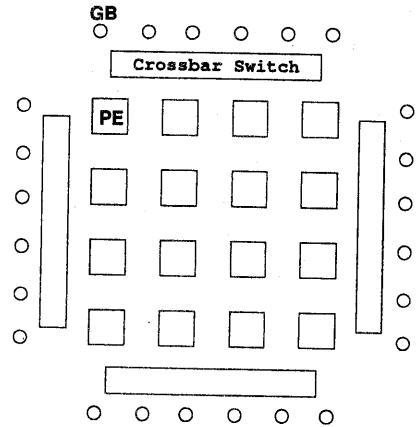


図 6: ウェハ上の PE と SW の配置

から、式 (6) は以下のように書き換えられる。

$$W = \sqrt{m}W_{PE} + 2(W_{SW} + W_{GB}) \quad (7)$$

$A = W^2$  であるから、式 (7) を 2 乗してレイアウト面積  $A$  は次式で表せる。

$$A = mW_{PE}^2 + 4(W_{SW} + W_{GB})\sqrt{m}W_{PE} + 4(W_{SW} + W_{GB})^2, \\ = mA_{PE} + 4A_{SW} + 4\sqrt{m}W_{PE}(W_{SW} + W_{GB}). \quad (8)$$

#### 4.2 スイッチの入出力数とレイアウト面積

SW はクロスバスイッチを用いているので、SW の入出力に大きく依存する。そこで SW の入出力数を求める。SW には SW と同じウェハ上から異なるウェハ間結合のための結線と、他のウェハからの結線が入出

力として必要となる。よって SW の入出力数は以下のようになる。

$$N_{SW} = W_{SW} + W_{GB} = \frac{\sqrt{m}}{2} + \frac{C_{peak}}{4} \quad (9)$$

式(9)から SW は  $m$  と  $C_{peak}$  のパラメータでその大きさがほぼ決まってくる。ここで WSI へ実装する超並列計算機を考えた場合、パラメータ  $m$  は  $C_{peak}$  に比べてかなり小さいので SW は  $C_{peak}$  にのみ依存すると言ってよい。また、SW のアービタ等、周辺のハードウェアには交点スイッチの全体とほぼ等しいだけのゲート数が必要であると仮定すると、 $C_{peak}$  に掛かる入出力以外の配線面積は  $C_{peak}$  の 2 乗の 2 倍によって全て隠蔽される。よって  $A_{SW}$  は、 $C_{peak}$  と  $A_{GB}$  から以下のようになる。

$$A_{SW} = C_{peak}^2 A_{GB} / 8 \quad (10)$$

式(8)(10)よりレイアウト面積  $A$  は、以下のように簡略できる。

$$A = m A_{PE} + \frac{1}{2} C_{peak}^2 A_{GB} + C_{peak} \sqrt{2m A_{PE} A_{GB}} \quad (11)$$

## 5 従来ネットワークとの性能比較

### 5.1 直径とコスト

BCT と従来ネットワークを直径、コストについて比較する。比較する従来ネットワークは階層型相互結合網である dB Cube [6], CCC [7], Hyper dB [8] とし、BCT については双方向とする。図 7 に各ネットワークの直径とノード数の関係を示す。また図 8 に各ネットワークのコストとノード数の関係を示す。図 7, 図 8 から、BCT は dB Cube と比較すると直径、コストに優れている。dB Cube は、ネットワークのサイズによりリンク次数が増加するため BCT の方がより実装向きである。CCC と比較すると同程度の直径を示す。実装の面から考えると、CCC は、上位階層に Hyper Cube を用いており、大規模ネットワークにおいて物理リンクが多くなるため上位レベルに少ない物理リンクで結合できる de Bruijn を用いた BCT の方が簡単である。Hyper-dB と比較すると、Hyper-dB より直径が長い、Hyper-dB は、ネットワークサイズに応じてリンク次数が極めて大きくなるので、リンク次数、コストの点で Hyper-dB より遥かに優れていることが分かる。

### 5.2 レイアウト面積

レイアウト面積の比較を式(11)と以下の条件で従来の相互結合網と比較する [5]。

1. PE はゴールドバンプの 32 倍の大きさとし、 $A_{PE} = 32$ ,  $A_{GB} = 1$  とする。

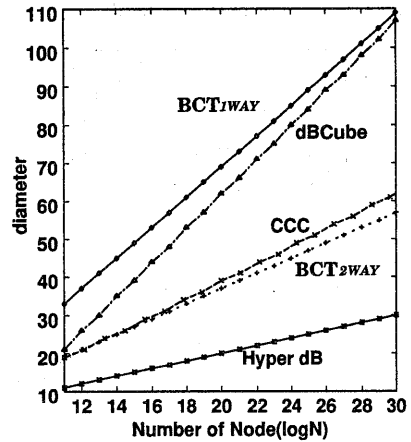


図 7: 各ネットワークの直径とノード数の関係

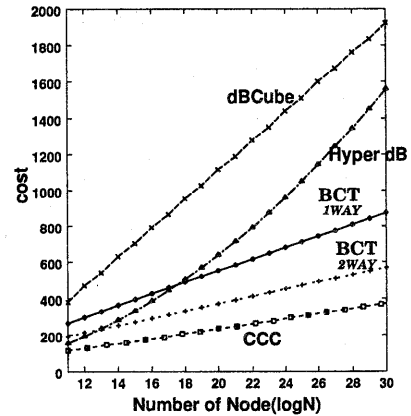


図 8: 各ネットワークのコストとノード数の関係

2. 全体の PE 数を  $4096(2^{12})$ , ウェハ一枚当たりの PE 数を  $256(2^8)$  とする。
3. 比較する相互結合網は、2DTorus, 3DTorus, Hyper Cube, BCT とする。

以上の条件での  $C_{peak}$  を表 4 に示す。

図 9 に 2DTorus を 1 とした各ネットワークの正規化レイアウト面積を示す。BCT は 3DTorus や Hyper Cube に比べ、非常に小さく構成できることが分かる。BCT はウェハ間結線数  $C_{peak}$  を非常に小さくできるために従来の相互結合網と比べて十分小さいレイアウト面積でネットワークを構成できる。

表 4:  $C_{peak}$  の比較

	$C_{peak}$
2DTorus	160
3DTorus	512
Hyper Cube	2048
BCT( $q=0, L=2$ )	58

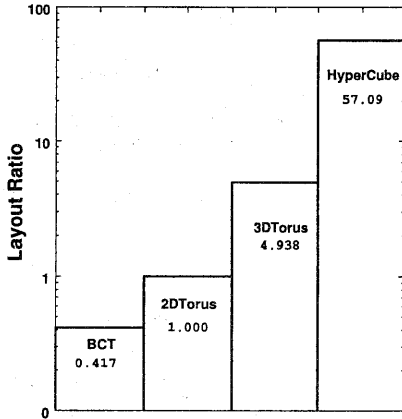


図 9: 各ネットワークの正規化レイアウト面積

## 6 まとめ

本論文では、2DTorus と de Bruijn ネットワークを用いた超並列向き階層型相互結合網 BCT を提案した。BCT は、ウェアスタック構造実装を考慮した階層間リンク数の少ないネットワークである。また、階層数が BM の大きさに制限されるものの、多階層構成が可能である利点を有している。BCT のネットワーク性能を理論的に求め従来の階層型相互結合網と比較した。その結果、階層間のリンク数を減らしたにも関わらず、従来の階層型相互結合網とほぼ同程度の性能があることが分かった。またレイアウト面積について考察し、BCT が従来の相互結合網より小さいレイアウト面積で構成することを示した。

今後の課題は、BCT の適応型ルーティングを提案し、動的通信性能について詳しく検討することである。

謝辞 本研究は文部省科学研究助成を用いて行なわれた。関係者各位に深謝する。

## 参考文献

- [1] 堀口 進, “ウェハ規模超密度集積回路について,” Hybrid. vol.6, No.1, pp16-21, 1990
- [2] Y.R.Potlapalli, “Trend in Interconnection Networks Topologies: Hierarchical Networks,” Int'l. Conf. on Parallel Processing Workshop, pp24-29, 1995
- [3] John Carson, “The emergence of stacked 3D Silicon and its impact on microelectronics systems integration,” Proc. 1996 Int'l. Conf. on Innovative Systems in Silicon, pp.1-8, Oct 1996
- [4] S.Okugawa, “Characteristics of the de Bruijn Network for Massively Parallel Computers,” Trans. IEICE, vol. J75-D-I, no8, pp592-599, Aug 1992
- [5] K.D.Gann, “Neo-Wafer 3D Packaging,” IMAPS 3D Packaging Advanced Technology Workshop, Nov 1998
- [6] C.Chen, D.P.Agrawal and J.R.Burke, “dBCube: A new class of hierarchical multiprocessor networks and its area efficient layout,” IEEE Trans. Para.&Dist. Sys. Vol.4, No.12, pp.1332-1343, Dec. 1993.
- [7] F.P.Preparata and J.E.Vuillemin, “The cube connected cycles: A versatile network for parallel computation,” Commun. ACM, Vol.24, No.5, pp. 300-309, May 1981
- [8] E.Granesan and D.K.Pradhan, “The hyper-de Bruijn networks: Scalable versatile architecture,” IEEE Trans. Para.& Dist. Sys. Vol.4, No.9, pp.962-978, Sep, 1993.