Vol.2009-ARC-186 No.11
Vol.2009-HPC-123 No.11
2009/12/1

誘導結合を用いたフィールドスタッカブルCMPのための3-D NoCアーキテクチャの検討

松谷宏紀[†1,†2] 鯉渕道紘[†3]
黒田忠広[†4] 天野英晴[†4]

本研究では誘導結合によるチップの3次元積層技術に着目し，アプリケーションに応じて積層するチップの枚数や種類を変更可能なCMPアーキテクチャについて検討する．本研究で提案する3次元CMPでは，垂直方向の通信インタフェイスおよび任意の平面オンチップネットワーク（NoC）を持ったチップを積層するだけで，隣接チップ同士で経路情報の交換を行い，パッケージ全体として1つの3次元ネットワークを自動的に形成する．本論文では，このような3次元CMPのフルシステムシミュレーション環境を構築，水平NoCを持つチップと持たないチップを混載積層した3次元CMPをシミュレーションし，提案するCMPアーキテクチャが正しく動作することを確認する．さらに，SPLASH-2ベンチマークの実行結果，および，ハードウェアコストと消費エネルギーの予備評価を通して，3次元化によるメリットを示す．

A 3-D NoC Architecture for Field Stackable CMPs using Inductive Coupling

Hiroki Matsutani,[†1,†2] Michihiro Koibuchi,[†3]
Tadahiro Kuroda[†4] and Hideharu Amano[†4]

In this paper, we discuss a novel 3-D CMP architecture, in which the number and types of chips stacked in a package can be changed in response to the applications running on the CMP, by using the inductive coupling based 3-D IC technology. Each chip in the proposed 3-D CMP architecture has vertical communication interfaces and an arbitrary horizontal Network-on-Chip (NoC). By stacking such chips, their topology and routing information is automatically exchanged and a 3-D network across them is then formed. In this paper, we develop a full system simulation environment for the proposed CMP architecture. A heterogeneous 3-D CMP, in which some chips have their own horizontal NoCs while the others do not, is demonstrated in order to confirm the correct operation of the proposed system. As preliminary evaluations, we show the SPLASH-2 benchmark performance, hardware cost, and energy consumption of the proposed CMP architecture so as to show its feasibility.

1. はじめに

半導体技術の微細化にともない1チップ上に複数のマイクロプロセッサを実装できるようになった．コンシューマ用途においても2コアや4コアの製品が広く普及しており，コアの数は今後も増え続けると予想される．コンシューマ用途のマルチコアでは，プログラミングの容易さから，すべてのコアが同一のメモリ空間を共有する共有メモリ型のチップマルチプロセッサ（CMP）が現実的と言える．ただし，複数のプロセッサが単一のキャッシュを共有するため，キャッシュアクセスに十分な帯域を確保しないとプロセッサ数に見合った性能向上は期待できない．そこで，キャッシュを複数のキャッシュバンクに分割して帯域を稼ぐアーキテクチャ（Non-Uniform Cache Architecture，NUCA）[1),2)]が有望視されている．NUCAではプロセッサおよびキャッシュバンクをNetwork-on-Chip（NoC）[3)]で接続し，データ転送はオンチップルータを介したパケット転送によって行う．

このようなCMP内のネットワーク化によって，理論上，多数のプロセッサやキャッシュバンクを統合できるようになった．ところが，このような大規模CMPを効率良く利用するには依然として以下の問題を解決しなければならない．

- **問題点1：** プロセッサ，キャッシュバンク，ルータ利用率の不均衡化：
  アプリケーションには並列化できる部分とシングルプロセッサで逐次実行しなければならない部分がある．アプリケーションによっては，逐次実行部分に律速されてプロセッサ数を増やしてもそれに見合った性能向上が得られない場合がある．また，NUCAではプロセッサとキャッシュバンクの距離が均一ではなく，アクセスするメモリブロックによって通信遅延が大幅に変化したり，トラフィックに偏りが生じる．

†1 東京大学大学院 情報理工学系研究科
Graduate School of Information Science and Technology, The University of Tokyo
†2 日本学術振興会特別研究員 (SPD)
Research Fellow of the Japan Society for the Promotion of Science (SPD)
†3 国立情報学研究所 / 総合研究大学院大学
National Institute of Informatics / The Graduate University for Advanced Studies
†4 慶應義塾大学 理工学部
Faculty of Science and Technology, Keio University

Vol.2009-ARC-186 No.11
Vol.2009-HPC-123 No.11
2009/12/1

- **問題点 2：** ルータホップ数の増加にともなう NoC の消費電力の増加：
  コアの数が増加すれば，その分，ルータの数が増え，パケット転送のたびに経由しなければならないルータの数（ホップ数）が増える．ホップ数に応じてルータおよび配線リンクの消費エネルギーがリニアに増える．

問題点 1 を解決するには，アプリケーションに応じて，プロセッサやキャッシュバンクの数，および，それらの接続関係を柔軟に変更可能にする必要がある．問題点 2 に関しては，すでにさまざまな低消費電力技術が提案されている[4),5)] ものの，最もドラスティックな改善案はチップの 3 次元積層であると言える．2 次元トポロジよりも 3 次元トポロジのほうがホップ数が小さく，期待されるスループット性能も高い．また，配線遅延やその消費エネルギーが問題になっている昨今，mm オーダの水平リンクを数十 μm オーダの垂直リンクに置き換えることはメリットが大きい．

本研究では，上記の問題を同時に解決するために，近年，実用化に向けて急速に研究開発が進んでいる誘導結合によるチップの 3 次元積層技術を用いる．本論文では，そのための第一段階として，誘導結合による 3 次元 CMP アーキテクチャについて検討し，アプリケーションの実行時間，結合網の面積，消費エネルギーに関する予備評価を示す．

本論文の構成は以下のとおりである．まず，2 章で，近年盛んに研究されている共有メモリ型の CMP アーキテクチャとその NoC について述べる．3 章でフィールドスタッカブル CMP について説明し，4 章でそのための 3-D NoC アーキテクチャについて検討する．5 章で予備評価を示し，6 章で本論文をまとめる．

2. 共有メモリ型 CMP アーキテクチャ

本章では，近年盛んに研究されている 2 次元の CMP アーキテクチャとその NoC について述べる．

図 1 に本論文でベースラインとする 2 次元の共有メモリ型 CMP のチップレイアウトを示す．これは文献 2) で紹介されている「2010 年の CMP」をもとに，L2 キャッシュバンクの数など一部パラメータを修正したものである．

図に示すようにチップ内にプロセッサコア（CPU）を 8 個持つ．各プロセッサコアは非共有の L1 データキャッシュ（L1 D$），L1 命令キャッシュ（L1 I$）を持つ．L2 キャッシュ（L2$）はすべてのプロセッサ間で共有し，token coherence protocol[6)] によるコヒーレンス制御を行う．キャッシュアクセスを高速化するため，キャッシュの構成は SNUCA（statically mapped, non-uniform cache architecture）[1)] とする．具体的には，L2 キャッシュを多数のキャッシュバンクに分割し，ブロックインデックスの下位数ビットをもとに割り当てるキャッシュバンクを決める[⋆1]．メインメモリおよびディレクトリコントローラ（Dir）はチップ外にあると仮定する．

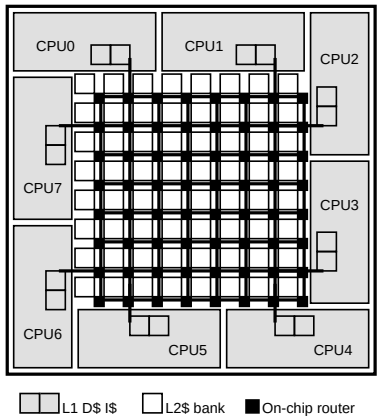

図 1 ベースラインとする 2 次元の CMP.

ここでは，プロセッサ（ローカル L1$ を含む）と L2$バンクを接続するために，オンチップネットワークを用いる．図 1 の例では，黒い四角がオンチップルータであり，オンチップルータが 8 × 8 の 2 次元メッシュ状に相互接続されている．パケットルーティングとして，メッシュにおいて最もシンプルかつ一般的な次元順ルーティングを用いる．

以降の章では，このような CMP アーキテクチャを 3 次元化することを考える．

3. 誘導結合による 3 次元 CMP

本章では，誘導結合を用いた 3 次元 CMP（フィールドスタッカブル CMP）のアーキテクチャについて議論する．

3.1 誘導結合による 3 次元積層

チップもしくはウェハの 3 次元積層技術として，これまでに様々な技術が実用化されおり，とりわけ，1) マイクロバンプ[7),8)]，2) 貫通ビア（Through-silicon via，TSV）[9),10)]，3) 容

⋆1 頻繁に使われるキャッシュブロックをプロセッサの近隣に動的に移動させることもできる（DNUCA，dynamically mapped, non-uniform cache architecture[1)]）．しかし，CMP では，あるプロセッサの近くにあるバンクは別のプロセッサからは遠くなってしまうため，結果的に高い効果は期待できないと言われている[2)]．

Vol.2009-ARC-186 No.11
Vol.2009-HPC-123 No.11
2009/12/1

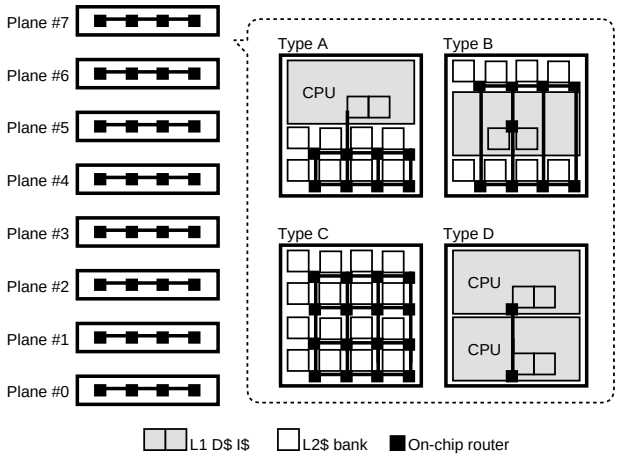

図 2 3 次元 CMP のコンセプト.

量結合[11]，4) 誘導結合[10),12),13)] などが代表的である．

その中でも誘導結合による 3 次元積層には次の特徴がある．

- 誘導結合は非接触型（ワイヤレス）である．接触型のマイクロバンプや貫通ビアと比べ，積層するチップの種類や枚数を柔軟に変更できる．
- 容量結合も非接触型であるが，2 枚のチップを face-to-face で接続するため積層枚数は 2 枚に限られる．一方，誘導結合は積層枚数に制限はなく，場合によっては，複数チップに対しデータをマルチキャストすることもできる．
- 2007 年に発表された 90nm プロセスのデータにおいて，データ転送エネルギー 0.14pJ/bit，チャネルサイズ 30μm×30μm，データ転送レート 1Gbps と高帯域・低消費エネルギーを実現している[13]．

1 章で述べたとおり，今後さらに大規模化するであろう CMP を効率的に利用するために，アプリケーションに応じて，プロセッサやキャッシュバンクの数，および，それらの接続関係を柔軟に変更可能にする必要がある．このような CMP を実現するために，本研究では，上記の特徴を兼ね備えた誘導結合によるチップの 3 次元積層技術に着目する．

3.2 フィールドスタッカブル CMP

図 2 に誘導結合による 3 次元 CMP のコンセプトを示す．この例では，図 1 で単一チップ上に実装されていた各種コアを 8 枚のプレーンに分割して，垂直方向に積層している．非接触型の 3 次元積層技術を用いることでプレーンの種類や枚数を柔軟に変更できる．図 2 には 4 種類のプレーン（Type A-D）が図示してあり，例えば，Type A はプロセッサとキャッシュバンクを持つプレーン，Type C はキャッシュバンクのみのプレーン，Type D はプロセッサのみのプレーンである．

本論文では CMP を想定しているが，このようなコンセプトは通常の SoC の置き換えとして広く応用できる．例えば，プロセッサ，メモリ，アナログ回路（センサ等）を別個の汎用チップとして調達し，それらを積み木のように組み合わせることで所望のシステムを構築できる．IP コア同士を組み合わせてマスクパターンを新規に作る SoC と異なり，出来合いの汎用チップを組み合わせる点でコスト的に有利である．実際に，我々は誘導結合を用いた動的再構成プロセッサのチップ試作を行い，実機で動作することも確認している[14]．

次章では，このようなフィールドスタッカブル CMP において求められる 3 次元オンチップネットワーク技術について議論する．

4. 3 次元 CMP のための 3-D NoC アーキテクチャ

本研究で想定するような非接触の積層技術を用いた CMP では，任意のプレーンを積層することでパッケージ全体として 1 つの 3 次元 NoC を形成する（図 2）．

4.1 ネットワークの動的な認識

各プレーンが持つ水平 NoC の形状は様々である．例えば，2 次元メッシュ状の NoC を持つプレーン，一部リンクが欠損した NoC を持つプレーン，いっさい NoC を持たないプレーンなどが考えられる．ハードウェアコストを削減するために通信量の少ない水平リンクを削除する場合，ハードマクロに邪魔されてルータ（リンク）を配置できない場合，製造時の故障によって一部の水平リンクが利用できない場合など要因は様々である．

水平 NoC を一部もしくはまったく持たないプレーンは，同一プレーン上のコアと通信する際に他のプレーンの水平 NoC を借りて通信する必要がある．つまり，3 次元 CMP を形成するには他のプレーンの経路情報が必要であり，CMP の起動時にプレーン同士で経路情報を交換し合い，自動的にネットワークを形成できなければならない．

本研究で実装したフィールドスタッカブル CMP シミュレータ（後述）では，ネットワークの起動時にダイクストラ法を用いて最短経路の探索を行う．具体的には，全体のトポロジ情報をもとに各ルータからすべての宛先ルータへの最短経路を計算，各ルータのルーティングテーブルを設定する．次節では水平 NoC を持つプレーンと持たないプレーンを混載させた 3-D NoC アーキテクチャを示すが，上記の方法を用いることで，このような不規則 3-D

Vol.2009-ARC-186 No.11
Vol.2009-HPC-123 No.11
2009/12/1

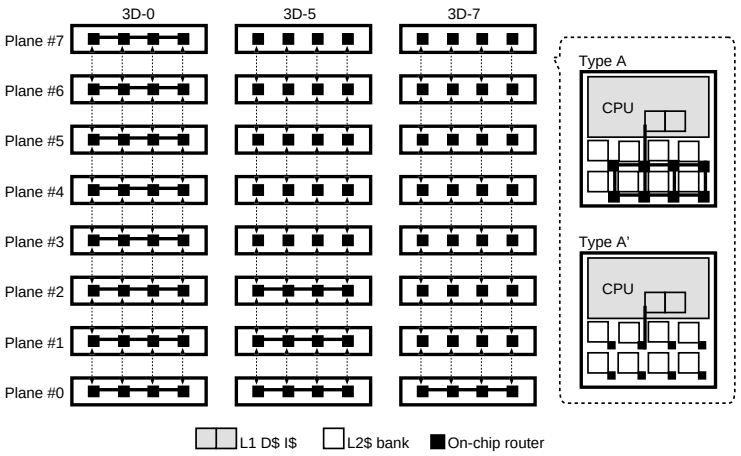

図 3 低コストな 3-D トポロジの例（図中の 3D-5 は 8 プレーン中 5 プレーンが平面 NoC を持たない（Type A'）という意味）.

NoC においても安定した通信を実現できる.

4.2 低コストな 3-D トポロジの設計

図 3 にヘテロジーニアスなチップ積層の例を示す．この例では水平 NoC を持つプレーン（Type A）と持たないプレーン（Type A'）を混載している．図中の 3D-0（マイナス 0）はすべてのプレーンが Type A であるが，3D-5（マイナス 5）では Type A は 3 個だけである．3D-7（マイナス 7）に至っては Type A は最下層のプレーン#0 のみであり，すべての水平方向の通信はプレーン#0 の NoC を経由して行うことになる.

つまり，最低限最下層のプレーンのみ水平 NoC を提供すれば，他のプレーンは垂直方向の通信インタフェイスさせ持てば水平 NoC がなくとも全体の接続性を確保できる．水平 NoC が少なければ少ないほど NoC のハードウェアコストは減るが，通信のたびに異なるプレーンの NoC を利用するため平均ホップ数および通信遅延が増え，アプリケーションの性能に影響が生じる．したがって，できるだけ利用率の低い水平リンクを削除することが望ましい．要求される性能とコストに応じて水平リンクを削除すれば，あとは上述のネットワーク自動認識プロトコルによって自動的に 3-D NoC が形成される.

次章では，3 次元 CMP の予備評価として 3D-n（$0 \leq n < 8$）のアプリケーション性能，ハードウェア量，消費エネルギーについて解析する.

表 1 CMP の評価パラメータ

Processor	UltraSPARC-III
L1 I-cache size	16 KB (line:64B)
L1 D-cache size	16 KB (line:64B)
# of processors	8
L1 cache latency	1 cycle
L2 cache size	256 KB (assoc:4)
# of L2 cache banks	64
L2 cache latency	6 cycle
Memory size	4 GB
Memory latency	160 cycle

表 2 NoC の評価パラメータ

Topology	8×8 mesh (2D) 2×4×8 mesh (3D)
Routing	dimension-order
Switching	wormhole
# of VCs	4
Buffer size	4 flit
Router pipeline	[RC][VSA][ST]
Flit size	128 bit
Control packet	1 flit
Data packet	5 flit

5. 予備評価

まず，本研究が対象としている 3 次元 CMP のフルシステムシミュレータについて述べる．次に，このシミュレータを用いて 2D NoC および 3D-n NoC を予備評価を示す.

5.1 評価環境

2 章〜4 章で述べた 2 次元および 3 次元 CMP をシミュレーションする．キャッシュアーキテクチャは SNUCA とし，キャッシュコヒーレンス制御には token coherence protocol を使用する．**表 1** にプロセッサとメモリシステムの詳細，**表 2** に NoC のパラメータを示す.

OS まで含めた CMP のフルシステムシミュレーションのために GEMS[15] および Simics[16] を組み合わせて使用する．CMP の NoC 部分には，GEMS に含まれているネットワークモデル Garnet[17] を使用した．今回は Garnet の detailed network model を改変することで，誘導結合による 3 次元 CMP をモデリングできるようにした.

ルータのパイプライン段数は 3 ステージ（RC，VSA，ST）とし，ルータが 1 フリット転送するのに 3 サイクルかかるものとした．加えて，フリットが水平リンク上を移動する（LT）ためにさらに 1 サイクルかかるものとした．誘導結合による垂直リンクのリンク遅延も 1 フリット当たり 1 サイクルとした.

アプリケーションとして SPLASH-2 ベンチマーク[18] から 12 種類のプログラムを用いた．8 コアの CMP を想定したシミュレーション環境で Sun Solaris 9 を動作させ，そのうえで Sun Studio 12 の開発環境を用いて 12 種類のプログラムをコンパイルした．個々のプログラムは，スレッド数を 16 として Solaris 9 上で動作させた．各プログラムの実行時間として，処理のコア部分の実行サイクル数をカウントした.

Vol.2009-ARC-186 No.11
Vol.2009-HPC-123 No.11
2009/12/1

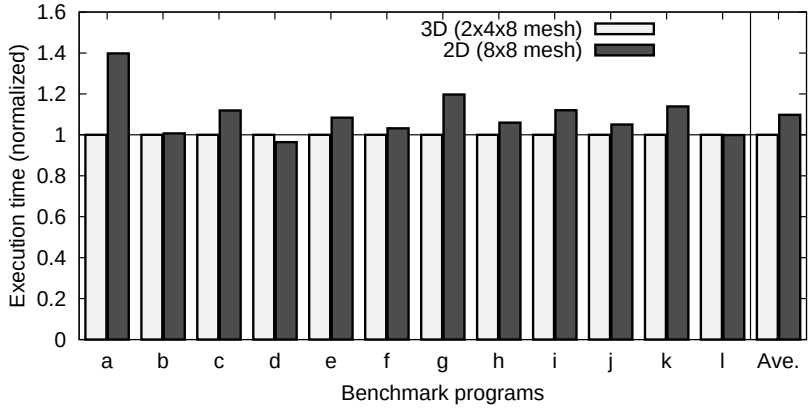

図 4 2 次元メッシュと 3 次元メッシュにおける SPLASH-2 ベンチマークの実行時間（3 次元メッシュの実行時間を 1 として正規化）．プログラムは (a) cholesky, (b) radix, (c) lu, (d) fft, (e) barnes, (f) radiosity, (g) ocean, (h) raytrace, (i) volrend, (j) water-nsquared, (k) water-spatial, (l) fmm の 12 種類.

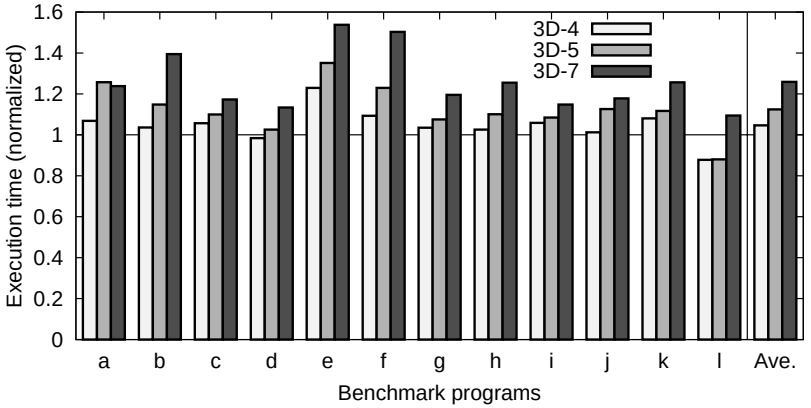

図 5 3 次元メッシュにおいて平面 NoC 数を変えたときの SPLASH-2 ベンチマークの実行時間（3D-5 は 8 プレーン中 3 プレーンが平面 NoC を持つ）．プログラムは (a) cholesky, (b) radix, (c) lu, (d) fft, (e) barnes, (f) radiosity, (g) ocean, (h) raytrace, (i) volrend, (j) water-nsquared, (k) water-spatial, (l) fmm の 12 種類.

5.2 アプリケーションの実行時間

まず，2D NoC と 3D NoC における SPLASH-2 ベンチマークの実行時間を比べる．ここでは 2D NoC は 8x8 メッシュ構成（図 1）で，3D NoC は 2x4x8 メッシュ構成（図 3 の 3D-0）である．シミュレーション結果を **図 4** に示す．このグラフでは 3 次元メッシュにおける実行時間を 1 として正規化してある．2 次元メッシュより 3 次元メッシュのほうが平均ホップ数が小さく帯域も広い．そのため，2 次元メッシュにおけるアプリケーションの実行時間は 3 次元メッシュのそれと比べて平均で 9.8% 長くなった．

次に 3D NoC-n における SPLASH-2 ベンチマークの実行時間について見ていく．ここでは $0 \leq n < 8$ としてシミュレーションしたが，紙面の都合から**図 5** では 3D-4，3D-5，3D-7 の結果のみ示す．このグラフでは 3D-0 における実行時間を 1 として正規化してある．3D-4，3D-5，3D-7 の順に内包する水平 NoC の数が減るためハードウェアコストが小さくなるが，その分だけ通信ホップ数，通信遅延，ネットワーク帯域が犠牲になる．その結果，3D-0 における実行時間と比較して，3D-4 は平均 4.7% 増，3D-5 は平均 12.5% 増，3D-7 は平均 25.9% 増となった．2D NoC の実行時間は 3D-0 と比較して 9.8% 長いことを考えると，3D-4 のように水平 NoC の数を半分にしたとしても 2D より 3D のほうが性能が良いという結果になった．

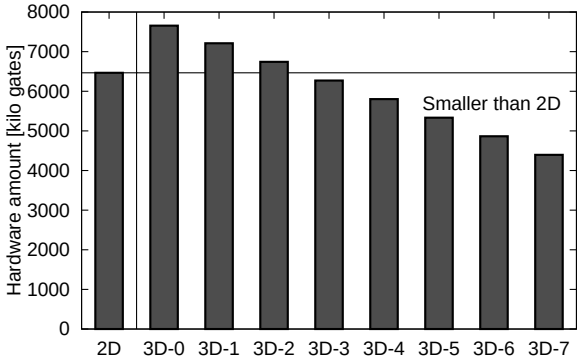

図 6 3 次元メッシュにおいて平面 NoC 数を変えたときの結合網の面積（インダクタの面積は含まず）.

次節では 3D-4 のように水平 NoC の数を減らすことでどれだけハードウェア量を削減できるか見積もる．

5.3 オンチップルータのハードウェア量

2D NoC と 3D-n NoC におけるオンチップルータのハードウェア量を見積もる．これら

Vol.2009-ARC-186 No.11
Vol.2009-HPC-123 No.11
2009/12/1

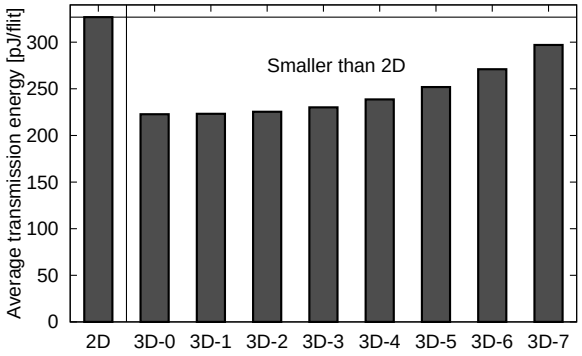

図 7 3 次元メッシュにおいて平面 NoC 数を変えたときの平均フリット転送エネルギー（インダクタのエネルギーは含まず）.

のトポロジではルータが合計 64 個使われているが，各ルータの次数（物理ポート数）はトポロジごとに異なる．例えば，2 次元トポロジより 3 次元トポロジのほうが垂直方向のチャネルがある分，ルータの次数が大きい．そこで，トポロジごとに次数 p のルータ（$2 \leq p < 8$）の個数をカウントし，次数 p ルータのゲート数と掛け合わせることで NoC のトータルのハードウェア量を計算した．

ルータのゲート数を見積もるために，文献 19) で使用したオンチップルータを 表 2 の仕様に合うように改造した．コヒーレンスプロトコルによる end-to-end のプロトコルデッドロックを防ぐために仮想チャネル数は 4 本とした．Verilog-HDL で記述された上記の次数 p ルータを Synopsys Design Compiler を用いて合成し，2 入力 NAND 換算でゲート数を求めた．ライブラリとしてコア電圧 1.20V の 65nm CMOS プロセスを用いた．

図 6 に 2D NoC と 3D-n NoC のトータルのゲート数を示す．2D と比べ 3D-0 のほうが垂直方向のチャネルを持つ分だけハードウェア量が大きい．一方，3D-n では n が大きくなる（水平 NoC の数が減る）にしたがいコストが減っていき，$n \geq 3$ で 2D よりもコストが小さくなる．前節で示したとおり，3D-3 よりさらにコストの小さい 3D-4 ですら 2D より性能が高いため，3 次元化によるメリットは大きいと言える．

5.4 フリット転送エネルギー

最後に，2D NoC と 3D-n NoC における平均フリット転送エネルギーを見積もる．平均フリット転送エネルギー E_{flit} は送信元から宛先まで w-bit のフリットを転送するのに要す平均エネルギーであり，次式を用いて計算される[20]．

$$E_{flit} = wH_{ave}E_{link} + w(H_{ave} + 1)E_{sw} \quad (1)$$

ただし，H_{ave} は平均ホップ数，E_{link} はリンク上を 1-bit データを転送するのに要すエネルギー，E_{sw} はルータが 1-bit データをスイッチングするのに要すエネルギーである．

E_{sw} はオンチップルータ回路の post-layout シミュレーションから算出した．具体的には，1) ルータ回路を Synopsys Astro で配置配線し，2) 実際のトラフィックを想定した post-layout シミュレーションにより配置配線後ネットリストの活性化率を求め，3) この活性化率をもとに Synopsys Power Compiler を用いて E_{sw} を計算した．

E_{link} は配線容量，リンクの配線長（ルータ-ルータ間距離），および，供給電圧によって決まる．ここでは，コア電圧 1.20V の 65nm プロセスにおいて，オンチップリンクに 1.0mm の semi-global 配線を用いるものとして E_{link} を計算した．なお，水平方向の移動距離に比べて，誘導結合による垂直方向の通信距離は数十 μm オーダと短いため，垂直方向の E_{link} は計算に含めていない．

図 7 に 2D NoC と 3D-n NoC の E_{flit} を示す．2D NoC において配線リンクで消費されるエネルギーの多くが 3D では垂直リンクに置き換わっているため 3D のほうが消費エネルギーが小さい．3D-n NoC は n が大きくなるにしたがい垂直方向の移動が増えてルータのスイッチング回数と E_{flit} が増加しているが，それでも，2D と比べてエネルギー効率の点で有利であると言える．

6. まとめと今後の課題

本研究で提案した 3 次元 CMP では，垂直方向の通信インタフェイスおよび任意の平面 NoC を持ったチップを積層するだけで，隣接チップ同士で経路情報の交換を行い，パッケージ全体として 1 つの 3 次元ネットワークを自動的に形成する．

本論文では，フルシステムシミュレータおよび実際のオンチップルータ回路を用いて，n 枚の水平 NoC を省略した 3 次元メッシュ（3D-n）の動作確認，および，予備評価を行った．その結果，1) 3D-4 のように半数の平面 NoC を省略した 3 次元メッシュでさえ，2 次元メッシュより高い性能を実現できた．2) 一般的に 3 次元メッシュは 2 次元メッシュよりコストが大きいが，3D-3 ないし 3D-4 では平面 NoC を一部省略したことで 2 次元メッシュよりもルータハードウェア量が小さくなった．また，3) CMP の 3 次元化によってフリットの転送エネルギーの点でも有利になった．

今回の予備評価では，議論を簡単にするため，誘導結合のためのインダクタ面積や消費エ

Vol.2009-ARC-186 No.11
Vol.2009-HPC-123 No.11
2009/12/1

ネルギーについては考慮しなかった．今後は，インダクタの詳細な回路パラメータを用いて性能，面積，消費エネルギーについて詳細に評価する予定である．さらに，本シミュレータ上でアプリケーションの特性を解析し，アプリケーションごとに積層するプレーンの種類や枚数をカスタマイズできるようにもする．

謝 辞 本研究は東京大学大規模集積システム設計教育研究センターを通し，シノプシス株式会社・日本ケイデンス株式会社の協力で行われた．また，本研究は日本学術振興会特別研究員奨励費の助成を受けて行われた．

参 考 文 献

1) Kim, C., Burger, D. and Keckler, S.W.: An Adaptive, Non-Uniform Cache Structure for Wire-Delay Dominated On-Chip Caches, *Proceedings of the International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS'02)*, pp.211–222 (2002).
2) Beckmann, B.M. and Wood, D.A.: Managing Wire Delay in Large Chip-Multiprocessor Caches, *Proceedings of the International Symposium on Microarchitecture (MICRO'04)*, pp.319–330 (2004).
3) Dally, W.J. and Towles, B.: Route Packets, Not Wires: On-Chip Interconnection Networks, *Proceedings of the Design Automation Conference (DAC'01)*, pp.684–689 (2001).
4) Matsutani, H., Koibuchi, M., Wang, D. and Amano, H.: Run-Time Power Gating of On-Chip Routers Using Look-Ahead Routing, *Proceedings of the Asia and South Pacific Design Automation Conference (ASP-DAC'08)* , pp.55–60 (2008).
5) Matsutani, H., Koibuchi, M., Wang, D. and Amano, H.: Adding Slow-Silent Virtual Channels for Low-Power On-Chip Networks, *Proceedings of the International Symposium on Networks-on-Chip (NOCS'08)* , pp.23–32 (2008).
6) Martin, M. M.K., Hill, M.D. and Wood, D.A.: Token Coherence: Decoupling Performance and Correctness, *Proceedings of the International Symposium on Computer Architecture (ISCA'03)*, pp.182–193 (2003).
7) Kumagai, K., Yang, C., Goto, S., Ikenaga, T., Mabuchi, Y. and Yoshida, K.: System-in-Silicon Architecture and its application to an H.264/AVC motion estimation fort 1080HDTV, *Proceedings of the International Solid-State Circuits Conference (ISSCC'06)*, pp.430–431 (2006).
8) Black, B., Annavaram, M., Brekelbaum, N., DeVale, J., Jiang, L., Loh, G.H., McCaule, D., Morrow, P., Nelson, D.W., Pantuso, D., Reed, P., Rupley, J., Shankar, S., Shen, J.P. and Webb, C.: Die Stacking (3D) Microarchitecture, *Proceedings of the International Symposium on Microarchitecture (MICRO'06)*, pp.469–479 (2006).
9) Burns, J., McIlrath, L., Keast, C., Lewis, C., Loomis, A., Warner, K. and Wyatt, P.: Three-Dimensional Integrated Circuits for Low-Power High-Bandwidth Systems on a Chip, *Proceedings of the International Solid-State Circuits Conference (ISSCC'01)*, pp.268–269 (2001).
10) Davis, W.R., Wilson, J., Mick, S., Xu, J., Hua, H., Mineo, C., Sule, A.M., Steer, M. and Franzon, P.D.: Demystifying 3D ICs: The Pros and Cons of Going Vertical, *IEEE Design and Test of Computers*, Vol.22, No.6, pp.498–510 (2005).
11) Kanda, K., Antono, D.D., Ishida, K., Kawaguchi, H., Kuroda, T. and Sakurai, T.: 1.27-Gbps/pin, 3mW/pin Wireless Superconnect (WSC) Interface Scheme, *Proceedings of the International Solid-State Circuits Conference (ISSCC'03)*, pp.186–187 (2003).
12) Miura, N., Mizoguchi, D., Inoue, M., Niitsu, K., Nakagawa, Y., Tago, M., Fukaishi, M., Sakurai, T. and Kuroda, T.: A 1Tb/s 3W Inductive-Coupling Transceiver for Inter-Chip Clock and Data Link, *Proceedings of the International Solid-State Circuits Conference (ISSCC'06)*, pp.424–425 (2006).
13) Miura, N., Ishikuro, H., Sakurai, T. and Kuroda, T.: A 0.14pJ/b Inductive-Coupling Inter-Chip Data Transceiver with Digitally-Controlled Precise Pulse Shaping, *Proceedings of the International Solid-State Circuits Conference (ISSCC'07)*, pp.358–359 (2007).
14) Saito, S., Kohama, Y., Sugimori, Y., Hasegawa, Y., Matsutani, H., Sano, T., Kasuga, K., Yoshida, Y., Niitsu, K., Miura, N., Kuroda, T. and Amano, H.: MuCCRA-Cube: a 3D Dynamically Reconfigurable Processor with Inductive-Coupling Link, *Proceedings of the Field-Programmable Logic and Applications (FPL'09)*, pp.6–11 (2009).
15) Martin, M. M.K., Sorin, D.J., Beckmann, B.M., Marty, M.R., Xu, M., Alameldeen, A.R., Moore, K.E., Hill, M.D. and Wood, D.A.: Multifacet General Execution-driven Multiprocessor Simulator (GEMS) Toolset, *ACM SIGARCH Computer Architecture News (CAN'05)*, Vol.33, No.4, pp.92–99 (2005).
16) Magnusson, P.S. et al.: Simics: A Full System Simulation Platform, *IEEE Computer*, Vol.35, No.2, pp.50–58 (2002).
17) Agarwal, N., Peh, L.-S. and Jha, N.: Garnet: A Detailed Interconnection Network Model inside a Full-system Simulation Framework, Technical Report CE-P08-001, Princeton University (2008).
18) Woo, S.C., Ohara, M., Torrie, E., Singh, J.P. and Gupta, A.: SPLASH-2 Programs: Characterization and Methodological Considerations, *Proceedings of the International Symposium on Computer Architecture (ISCA'95)*, pp.24–36 (1995).
19) Matsutani, H., Koibuchi, M., Amano, H. and Yoshinaga, T.: Prediction Router: Yet Another Low Latency On-Chip Router Architecture, *Proceedings of the International Symposium on High-Performance Computer Architecture (HPCA'09)* , pp.367–378 (2009).
20) Matsutani, H., Koibuchi, M., Hsu, D.F. and Amano, H.: Fat H-Tree: A Cost-Efficient Tree-Based On-Chip Network, *IEEE Transactions on Parallel and Distributed Systems*, Vol.20, No.8, pp.1126–1141 (2009).