主　論　文　要　旨

報告番号	甲　乙　第　　号	氏　名	西川　由理

主 論 文 題 目 ：

A Study of Interconnection Network for Many-Core Processors

Based on Traffic Analysis

（トラフィック解析に基づいたメニーコア型プロセッサの接続方式に関する研究）

（内容の要旨）

半導体プロセス技術の進歩により，単一チップ上にプロセッサやメモリ，I/O など複数の設計モジュールを搭載できるようになり，数十から数百個の演算コアを持つメニーコア型のプロセッサが登場している．この中でも，計算コアやキャッシュなどを格子状に配置したタイルプロセッサが次世代のプロセッサアーキテクチャとして有望視される傍ら，データ並列性の高いマルチメディア処理の効率化を図るための GPU をはじめとする SIMD アクセラレータの開発も盛んである．これらの背景を受け，今後のプロセッサアーキテクチャは NUCA 型タイルプロセッサと SIMD アクセラレータを単一チップ上に搭載する複合型となる可能性が高まっている．

タイル間，SIMD アクセラレータにおけるコア間，およびそれらのコンポーネント間の接続方式として，バスやチップ内ネットワーク（Network-on-Chip: NoC）を用いる手法が多数研究されている． 一方，複合型プロセッサにおいては，タイルプロセッサやアクセラレータ間，メモリ間等におけるバーストトラフィックのみならず，キャッシュコヒーレンシプロトコルのためのシグナルのような短いメッセージ長のトラフィックも増大することが予想される．このような異なる性質を持つトラフィックに対応するため，ヘテロジニアス型プロセッサは複数の NoC およびグローバル配線を併せ持つハイブリッドネットワークが有望である．本研究では，まず，このようなトラフィックの性質を明らかにした上で，NUCA 型タイルプロセッサと SIMD アクセラレータから成るヘテロジニアス型プロセッサにおける NoC と共有バス接続方式に関して，トラフィック解析に基づいた（1）ルーティング方式，および（2）ネットワークの性能モデルの提案を行う．

前者について，タイルプロセッサの結合網に適用可能な，短いトラフィック向けのルーティングである Semi-deflection routing を提案する．本手法は，単フリット型のパケットを想定した，仮想チャネルを用いない非最短完全適応型ルーティングである．提案手法により，従来の固定型および適応型ルーティングに比べ，トラフィックに偏りのある場合に 3.17 倍のスループット向上を達成し，wormhole ルーティング向けのルータに比べてハードウェア量も小さい．

後者について，SIMD 型メニーコアアクセラレータにおける共有バスと一次元接続網におけるレイテンシおよびスループットのモデル化を行う．具体的には，軽量な SIMD 型プロセッサである ClearSpeed 社の CSX600 を用いてトラフィック解析を行い，搭載 PE 数，転送データサイズ，データアライメントの有無から通信時間を見積もることのできるモデルを導出する．また，並列アプリケーションを用いた実際のデータ転送時間が，アプリケーショントレースと提案モデルを用いた予測通信時間に収まることを確認し，モデルを検証する．また，このモデルを用いて，共有バスのスループットを得るために必要な PE 数や，バス接続方式の有効性等について検討を行う．

上記のルーティングの性能評価および性能モデルから導かれることを踏まえ，最後に，複合型プロセッサにおけるハイブリッド型接続網について考察し，結論と将来の指針について述べる．

SUMMARY OF Ph.D. DISSERTATION

School of Science for Open and Environmental Systems	Student Identification Number	SURNAME, First name NISHIKAWA, Yuri

Title

A Study of Interconnection Network for Many-Core Processors Based on Traffic Analysis

Abstract

Improvements in VLSI technology has enabled integration of an increasing number of logic blocks such as processor cores, caches and I/O modules on a single chip, and multi- to many-core structure has become the mainstream of processor architectures. Currently, tile processor with non-uniform cache architecture (NUCA) is regarded as a new form of a many-core processor, while development of SIMD accelerators such as graphic processing units (GPUs) to deal with increasing data-parallel multimedia requirements. Thus, viable future processor architecture is a combination of NUCA-based tile processor and SIMD accelerators integrated into a single chip.

To connect both internal and external components of tiles and accelerators, various interconnection methodologies including shared buses and network-on-chips (NoCs) have been studied. One concern of such heterogeneous processor is an increase of transactions with various traffic characteristics such as burst point-to-point traffic between certain processing core and GPU cores, or cache coherence protocol signals with short message sizes. To address such characteristic diversities, a hybrid interconnect with multiple NoCs and global buses is one solution in order to meet latency-oriented and bandwidth-oriented traffic requirements. For the above aim, this thesis studies interconnection network including NoCs and shared buses for tile architectures and SIMD accelerators in terms of packet routing and network modeling.

As a packet routing exploration, a non-minimal fully- adaptive routing mechanism for single-flit-structured packet transfers called “Semi-deflection” is proposed. Semi-deflection routing guarantees deadlock-free packet transfer without use of virtual channels by allowing non-blocking transfer between specific pairs of routers. Evaluation results show that a router that supports Semi-deflection routing is almost equal to the hardware amount compared with those for existing deterministic and adaptive routing. As the result of throughput evaluation, Semi-deflection routing provides 3.17 times higher throughput.

As network analysis explorations, end-to-end latency and throughput of a simple shared-bus and a one-dimensional ring structure for SIMD many-core processors is analyzed and modeled in order to study their feasibility toward increasing number of processor cores. In concrete, we analyze Swazzle and ClearConnect, which are a one-dimensional NoC and a global bus for ClearSpeed’s CSX600, an SIMD processor consisting of 96 Processing Elements (PEs). By using the performance model derived from results of network analyses, we estimate the best- and the worst-case latencies for traffic patterns taken from several parallel application benchmarks, and confirm that actual latencies of the benchmarks fit in between the best- and the worst-case values.

Finally, we summarize our proposed routing technique for NUCA-based tile architectures and network analysis results of SIMD processors. Based on the above, we make a discussion on the structure of interconnection networks for many-core general-purpose tile processors, accelerators, and their combinations.