「情報社会を支える新しい高性能情報処理技術」
平成13年度採択研究代表者

中島　浩

（豊橋技術科学大学工学部　教授）

「超低電力化技術によるディペンダブルメガスケールコンピューティング」

1.研究実施の概要

本研究の目的は、100万プロセッサ規模のメガスケールコンピューティングによるペタフロップス計算を、実現性・信頼性・利用容易性のいずれにおいても現実的なものとするための基盤技術を確立し、かつ大規模プロトタイプを構築してその有効性を実証することにある。キーとなる技術は、(1) ハードウェア／ソフトウェア協調による低電力化技術、(2) 低コスト・ソフトウェア主導のディペンダブル技術、(3) グリッド／Peer-to-Peer (P2P) に基づくプログラミング技術、であり、これらに基づくプロセッサ、コンパイラ、ネットワーク、クラスタ構築、およびプログラミングの基盤技術の研究開発を行う。

本年度は各研究グループの担当要素技術と、グループ間にまたがるシステム構築技術について、以下のような主要な成果を得た。

(1) プロセッサグループ：SCMのリーク電流削減方式
(2) コンパイラグループ：プロトタイプMegaProtoの構築
(3) ネットワークグループ：RI2Nの高バンド幅・耐故障機能実装・評価
(4) クラスタ構築技術グループ：耐故障MPI・チェックポインティング最適化の実装
(5) プログラミング技術グループ：静的実行モデル生成機構の実装・評価

2.研究実施内容

【研究項目１：プロセッサグループ】

(1) リーク電流削減方式の検討

キャッシュメモリにおいては、利用されない領域を予測し、当該領域をリーク電流の少ないモードに切り替えてスタティック消費電力の削減を図る手法が提案されているが、予測が外れた場合には電力、実行時間の損失を被るという欠点がある。それに対し、提案するソフトウェア制御オンチップメモリを採用するアーキテクチャでは、データ転送をソフトウェアが制御するため、利用されない領域を完全に把握することが可能である。該当領域はデータを保持する必要がないため、SRAMセルへの電力供給を絶つことで、データを保持しない代わりにリーク電流を大きく削減できる回路技術であるGated-Vddを該当領域に対し実現する構成手法を検討した。また、

予備評価を行い、この手法により、性能を低下させることなく消費電力を効果的に削減できることを確認した。

(2) コンパイラによる低電力指向最適化

まず上記(1)に述べた、スタティック電力削減に必要となる、利用されない領域を完全に把握するアルゴリズムを開発し、昨年度までに開発したディレクティブベースコンパイラ上での実装に関して検討を加え、わずかな拡張で実現可能であることを確認した。次に、ソフトウェア制御オンチップメモリのうち、利用する領域を小さくすることでより積極的に上記(1)の手法を適用する方針について検討した。利用領域を小さくすればサイクル当たりのスタティック消費電力は減少するが、メモリ領域が小さくなるため、実行時間が増加する。この、実行時間とスタティック消費電力の間のトレードオフ関係を明らかにした。

(3) 低電力レジスタアーキテクチャ

性能を落とさずに低電力化を実現するレジスタ構成として、キャッシュと同様に多階層化し、小容量超高速レジスタファイルと大容量レジスタファイルからなる構成を提案し、その効果を検討した。頻繁にアクセスされるレジスタを小容量高速レジスタに効果的に割付けられる場合には性能的にも電力的にも有利だが、多くのプログラムにおいては、その割付アルゴリズムをさらに工夫しないとあまり効果がないこともわかった。

(4) FPGAを用いた評価環境の作成

FPGA(Vertex-II Pro)を用いた提案アーキテクチャ評価環境の試作を行った。プロセッサコアはPowerPC405ハードコアを利用し、FPGA内蔵block RAMを二つの領域に分割、それぞれのアクセスレイテンシを異なる値に設定した。片方を主記憶、片方をソフトウェア制御可能オンチップメモリ領域と見なすことで、提案アーキテクチャを疑似的に実現する。これにより、提案アーキテクチャの動作確認を、実ハードウェア上で確認することができた。また提案するアーキテクチャの特性を、FPGAを用いて定性的に評価することも可能となった。

【研究項目２：コンパイラグループ】

(1) プロトタイプMegaProto(#0)の構築

本研究の主要成果の一つである大規模プロトタイプMegaProtoの初期バージョンとして、1Uのマザーボードに16台のプロセッサ（TM5800, 933MHz, 256MBメモリ）とそれらを結合する２系統の1Gbイーサネットからなるクラスタユニット LP32を開発した。LP32のプロセッサあたりの消費電力は8Wと極めて小さく、またネットワーク等を含めたクラスタユニット全体での消費電力も300Wに抑えられている。この結果、ピーク演算性能・消費電力比は約50MFLOPS/Wという良好な値が得られた。また大きなネットワークバンド幅（ユニット内32Gbps, ユニット間16Gbps）を確保できているため、実効性能・消費電力比においても高い値が得られることが予想される。

(2) 低消費電力プロセッサによるクラスタシステムの構築と電力性能評価
上記のプロトタイプMegaProtoでの高性能・低電力計算の実現方式を詳細に検討するために、MegaProtoに用いられる低消費電力プロセッサCrusoeを用いて、４ノードからなるクラスタシステムを構築し、電力測定、性能評価を行った。その結果、複数の低消費電力プロセッサを用いることにより、いくつかのベンチマークプログラムにおいて、高性能プロセッサよりも電力あたりの性能を改善できることがわかった。特にメモリ最適化などの行った場合の並列プログラムでの電力削減効果が低消費電力プロセッサでは大きいことがわかった。ネットワークについても、電力消費特性の評価を進めるとともに、低消費電力プロセッサのDVS(Dynamic Voltage Scaling)機能を用いた場合の電力削減効果についても測定、考察をおこなった。

(3) アーキテクチャ駆動コンパイル、およびメモリ階層利用最適化アルゴリズム
コード生成部分について実装を進めるとともに、電力を削減するレジスタの構成の検討とともに必要な機能について再検討をしている。また、メモリ最適化アルゴリズムについてはアーキテクチャグループが利用できるようにアーキテクチャグループのディレクティブベースコンパイラへの組み込みを進めている。

(4) プロファイル取得機能の検討
プログラミング技術グループとともに、実行モデルの構築に必要なプロファイル、プロファイル取得機構のコンパイラの組み込みについて、検討した。

【研究項目３：ネットワークグループ】

(1) マルチリンクを利用した高バンド幅／耐故障性ネットワークの実装・評価
昨年度までのFast Ethernetベースのシステムに加え、dual-link Gigabit Ethernetを中心としたRI2N/USRの実装を行った。高バンド幅機能と耐故障機能を同時に実現する枠組みとして、マルチスレッド制御方式による動的リンク監視方式に基づく実装を行い、高バンド幅化におけるリンク資源の活用効率を上げると共に、リンク切断時の自動バックアップに対応する機構を組み入れた。また、RI2N/USRの性能評価を複数ノード間のping-poing転送を用いて行い、dual-linkによるバンド幅増強・動的負荷分散と、リンク切断時の動的バックアップ機能が動作することを確認した。

(2) RI2N/USRにおける故障検出機能
RI2N/USRにおいてリンク切断等の故障が発生した際、ソフトウェア的手法を用いて上位あるいは周辺ソフトウェアに対してこれを通知する機構を設計した。(1)に述べた実装において、これを検出・通知できることを確認した。また、オペレーティングシステムあるいは上位レイヤーソフトウェア、例えばクラスタ管理用ミドルウェアとの間でどのような通信を行うかについて検討した。

(3) MegaProto上におけるノード間ネットワーク構築法に関する検討
MegaProto上の16台の計算ノード間だけでなく、外部スイッチを介して複数のクラスタユニット間でのmulti-link通信を行う際のネットワーク構築法に関し、最大16ク

ラスタユニット（256計算ノード）に対するモデルを設計した。

【研究項目４：クラスタ構築技術グループ】

(1) モデルによるコスト予測システムの実装

PCクラスタ向け故障発生器の設計・開発を進めた。これは、本年度の計画であった故障発生時の並列プログラムの挙動を調査するために、PCクラスタ上に種々の故障を人為的に発生させるものである。本故障発生器は、パケットのロス、遅延、CPU負荷率の上昇、メモリ使用量の上昇等の故障を対象とする。今期はパケットをロスさせる発生器の実装・テストを行い、発生期のネットワーク対応の設計を進めた。パケットロスは、Linux カーネルのネットワークスタックに故障発生器のコードを挿入することで実現した。

(2) Omni/SCASH における動的負荷分散

昨年度までにメガスケールシステムを構成するクラスタ内のノード間性能の heterogeneity に対処するループ再分割機能を Omni/SCASH に対して実装した。本年度では、ループ再分割機能の評価の結果明らかになったデータローカリティ低下を改善する page fault counting に基づくページマイグレーション機能を実装した。ページマイグレーションによって Laplace を4ノードで実行した際におよそ60%の性能改善を得た。また profiled scheduling とページマイグレーションを組み合わせて使用した場合に、チューニングが不完全ながらstatic scheduling と同程度の性能を得られ、パフォーマンスロスなく動的スケジューリングが行える可能性を示した。また、page fault counting をベースにファーストタッチ、ネクストタッチなどと同等のページ配置手法を実験的に実装した。

(3) 耐故障性を備えたMPIの実現

昨年度プロトタイプを実装した Checkpointing/Restart を行う MPI に対し、Checkpointing のプロトコルを制御するデーモンを Coordinated Checkpointingのアルゴリズムを用いて実装した。ただし、この実装は現状、自動的にCheckpointingを行う部分までで、Restart/Migration に関してはまだ実行できない。また、耐故障性を付加する様々なコンポーネントを容易に可換にするため、共通のインターフェースを定義する必要がある。15年度は実験的にコンポーネント化のための最低限のインターフェースを定義、実装した。Coordinated Checkpointing の同期の時間を最大48CPU用いて測定した結果、24CPU以上になると、時間増加がゼロに近くなることを確認した。

(4) チェックポインティング最適化

並列Coordinatedチェックポイントを行なう際、チェックポイントファイルを保存する際のdiskI/Oの負荷が時間的に集中してしまう。これを解決する方法としてページ更新予測に基づく投機的チェックポイントを行うことにより、diskI/O負荷を分散させるSpeculativeチェックポイントを提案し、設計を行なった。また、このプロトタ

イプであるシングルプロセスのSpeculativeチェックポインタを実装した。これを使用し、擬似並列環境での実験を行なった。その結果、Speculativeチェックポイントを行なったものは、行なわないものと比べてチェックポイント時間が最大15%削減されることがわかり、Speculativeチェックポイントの有効性を確認した。

【研究項目５：プログラミング技術グループ】

(1) 実行モデル構築

MegaScriptにより記述されたメタプログラムを、静的には未知の変数を含む計算コスト式をノードとしたloop freeの制御フローグラフで表現した実行モデルに変換するトランスレータを構築した。また未知変数を具体化して実行コストを算出する機構を実装し、ＬＵ分解、モンテカルロ法、Mandelbrot集合、１次元波動方程式求解を対象にモデルの精度を評価した。その結果10%程度の誤差で実行性能を推定できることが明らかになり、モデル記述や構築手法の妥当性が確認された。

(2) 実行モデル精緻化

(1) で構築された実行モデルを、実行時情報を用いて精緻化する枠組みとして、タスクの特定部分の実行時間、ループ回転数、分岐成立回数、タスク中の特定変数の値などの収集機構を設計した。まず、メタプログラム中で収集すべきタスクの実行情報を指定するためのAPIを設計し、これらを抽出するためのメタプログラム・トランスレータの改造を行った。続いて【研究項目２】のコンパイラの機能を用いて、実行情報収集のためのコードをタスクプログラム中に埋め込むツールを作成した。

(3) 並列タスク実行機構

タスクスケジュールのための基本的なライブラリとして、タスクを静的に配置するスケジューラを構築した。また応用向けのライブラリとして、パラメータ探索ライブラリ、GA最適化ライブラリ、動的生成木探索ライブラリを設計・実装した。またこれらの開発過程で明らかになった、並列タスク実行機構のベースモジュールの問題点を解消するために、動的なタスクの生成・配置とそれに伴うタスク間ストリームの動的接続機能の設計・実装を行った。

(4) メガスケールシミュレータ

並列タスクを仮想的な大規模並列環境で実行して評価するためのツールとして、分散シミュレータAnastatiaの第１版を構築した。Anastatiaの実行メカニズムには、タスクプログラムをそのまま実行するエミュレーションモード、タスクをモデル化して実行するアブストラクションモード、エミュレーションの結果を利用してタスクモデルを構築するハイブリッドモードがあるが、第１版ではエミュレーションモードの実装を行った。モンテカルロ法とn-queenによる評価の結果、シミュレーションの精度は5%以下であり、分散実行により67～88%の高い並列化効率が得られた。

3.研究実施体制

プロセッサグループ

① 研究分担グループ長：中村　宏（東京大学先端科学技術研究センター、助教授）

② 研究項目：ソフトウェアとの協調最適化に基づく超低消費電力技術・高密度実装技術・高バンド幅技術

コンパイラグループ

① 研究分担グループ長：佐藤　三久（筑波大学計算物理学研究センター、教授）

② 研究項目：ハードウェアとの協調最適化に基づき低消費電力かつ高性能を実現するコンパイラ技術

ネットワークグループ

① 研究分担グループ長：朴　泰祐（筑波大学計算物理学研究センター、助教授）

② 研究項目：安価かつスケーラブルなディペンダブル高速ネットワーク技術

クラスタ構築技術グループ

① 研究分担グループ長：松岡　聡　（東京工業大学学術国際情報センター、教授）

② 研究項目：グリッド技術に基づくディペンダブルな大規模コモディティクラスタ構築技術

プログラミング技術グループ

① 研究分担グループ長：中島　浩　（豊橋技術科学大学情報工学系・教授）

② 研究項目：メガスケールかつディペンダブルなプログラミングモデル

4.主な研究成果の発表（論文発表および特許出願）

（1）論文（原著論文）発表

○ M. Kondo and H. Nakamura. Reducing Memory System Energy by Software-Controlled On-Chip Memory. IEICE Trans. Electoronics, Vol.E86-C, No.4, pp.550--588, Apr. 2003.

○ 高橋睦史，近藤正章，朴泰祐，高橋大介，中村宏，佐藤三久．HPC向けオンチップメモリプロセッサアーキテクチャSCIMAのSMP化の検討と性能評価．情報処理学会論文誌コンピューティングシステム，Vol.44, No. SIG6 (ACS 1), pp.76-86, May 2003.

○ Y. Sakae, S. Matsuoka, M. Sato and H. Harada. Preliminary Evaluation of Dynamic Load Balancing Using Loop Re-partitioning on Omni/SCASH. In Proc. Third IEEE/ACM Intl. Symp. Cluster Computing and the Grid / DSM, pp. 463--470, May 2003.

○ 丸山真佐夫，山本繁弘，大野和彦，中島浩．巻き戻し実行をサポートする並列プログラムデバッガ．SACSIS 2003, pp. 65-72, May 2003.

○ 大塚保紀，深野佑公，西里一史，大野和彦，中島浩．タスク並列スクリプト言語MegaScriptの構想．SACSIS 2003, pp. 73-76, May 2003.

○ 藤田元信，近藤正章，中村宏．ソフトウェア制御オンチップメモリ向け自動最適化コ

ンパイラの検討. SACSIS 2003, pp.173-174, May 2003.
○ 栄純明, 松岡聡, 佐藤三久, 原田浩. Omni/SCASH のループ再分割を用いた動的負荷分散拡張の実装と評価. SACSIS 2003, pp. 307--314, May 2003.
○ S. Miura, T. Boku, M. Sato and D. Takahashi. RI2N --- Interconnection network system for clusters with wide-bandwidth and fault-tolerancy based on multiple links. Proc. Intl. Symp. High Performance Computing 2004, LNCS-2858, pp.342-351, Oct. 2003.
○ 藤田元信, 近藤正章, 中村宏. ソフトウェア制御オンチップメモリ向け自動最適化コンパイラの提案. 情報処理学会論文誌コンピューティングシステム, Vol.44, No.SIG1(ACS4), pp.77--87, Jan. 2004.
○ M. Fujita, M. Kondo, and H. Nakamura. Data Movement Optimization for Software-Controlled On-Chip Memory. INTERACT-8, Feb. 2004.

（２）特許出願

該当なし