「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　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利用メモリ量最適化のための消費エネルギー理論式構築
(2) コンパイラグループ：プロトタイプMegaProtoの評価
(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システムCockoo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（研究目的、方法、結論などを記述）

【研究項目１：プロセッサグループ】

ソフトウェアとの協調最適化を実現するプロセッサアーキテクチャとして、チップ内のキャッシュメモリをアドレス指定可能なメモリ空間としても解放するアーキテクチャを提示し、コンパイラにより明示的にチップ内メモリを活用することで高性能低消費電力化を実現するアルゴリズムを開発した。消費電力に関してはスイッチング動作に伴う動的消費電力と、待機時にも消費される静的消費電力があり、半導体の微細化が進むと後者の影響が大きくなる。そこで、両方を考慮した、消費電力全体を低減し、しかも高性能を達成できるアルゴリズムの開発を目指した。

利用するチップ内メモリの増減が動的消費電力と静的消費電力に与える効果を、データの再利用性に富むプログラムである行列積と再利用性の乏しいベクトル内積で調べた。

その結果、利用するメモリ量を増やすと実行時間が短くなるため静的消費エネルギーは削減されるが、メモリ量自体が増えるため静的消費電力は増える。そのため静的消費エネルギーに関してはメモリ量との間でトレードオフの関係が成立することを確認した。また、利用するメモリ量を増やすとオフチップトラフィックが減少するため動的消費電力は減少することを確認した。上記の、利用するメモリ量と消費電力の間のトレードオフ関係は、定性的には２つのプログラムで共に成立するが、最適なメモリ量はプログラム毎に異なった。そこで、一般のプログラムに対し、最適な利用メモリ量を求められるコンパイルアルゴリズムの開発を目指し、与えられたプログラムとデバイスの特性から、メモリ量と消費エネルギーの関係を導出する理論式の構築を行った。

【研究項目２：コンパイラグループ】

① 実証システムMegaProtoに用いるCrusoeを用いた試作クラスタにおいて、並列環境におけるDVS制御ならびにキャッシュ最適化の電力削減効果等について電力測定を行い電力性能の評価を行った。この結果、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明らかにした。

② 実証システムMegaProtoの環境整備を行い、電力性能の評価を行った。また、電力状況をリアルタイムにモニターできるシステムを開発した。評価の結果、MegaProtoは既存の高性能プロセッサに比べて２倍以上の電力性能が得られることが実証された。またMegaProtoと電力モニターシステムを組み合わせたシステムをSC2004等にて展示し、本システムの有効性についてアピールした。

③ プロファイル駆動最適化のために、プロファイル取得機構の試作を行った。またプログラミング技術グループにおいて、Omniツールキットを用いてプロファイル取得・性能モデリングを実施していたが、これらを整理するとともに他の機能と統合についても検討を行った。

④ プロセッサグループのターゲットとしているMegaNodeプロセッサアーキテクチャ向けのコンパイラについては具体的にSH4をターゲットとして、コード生成の詳細設計、見直しを行った。

【研究項目３：ネットワークグループ】

① 単一スレッド下におけるマルチリンクによる高バンド幅対応のRI2N/USR-STの実装と評価

前年度までに実装したマルチスレッド版RI2N/USRを改良し、MegaProtoアーキテクチ

ャを意識したシングルスレッド版の高バンド幅対応のRI2N/USR-STを実装し性能評価を行った。システムライブラリの構成を大幅にシンプル化し、GbEthernetの単一リンク時に比べ、リンク２本を用い最大で1.67倍の性能が得られることを確認した。

② 耐故障機能と高バンド幅化機能を併せ持つRI2N/USRの実装と性能評価

マルチスレッド版RI2N/USRに耐故障機能を融合させ、これに合わせて高バンド幅化機能をチューニングし、ユーザレベルライブラリとして完全な機能を持つRI2N/USRを実装し、性能評価を行った。両機能を融合させたため、高バンド幅化の点では性能が落ちるものの、最大でリンク２本時に単一リンクの約1.4倍の性能が得られ、かつリンク故障後、これが復活した場合の高バンド幅機能が復元することを確認した。

③ MegaProto１号機におけるRI2N/USRの評価

MegaProtoの１号機のクラスタユニット内で、２本のGbEthernetリンクによるRI2N/USRのシングルスレッド版（高バンド幅機能のみ）及びマルチスレッド版（耐故障機能あり）を実装し性能評価を行った。現在のMegaProtoでは、次期２号機との連携上、PCIバスの性能が低いため単一リンク時での性能にそもそも問題があり、MegaProto２号機において本来の性能が発揮されるとの予測を立てた。

④ 上位階層との連携機能の概念設計

上位のクラスタ管理ソフトウェアにおける耐故障機能との連携に関し、RI2N/USR上でのマルチリンクの故障状況に応じてチェックポインティングの間隔を制御し、また上位管理ソフトウェアからのシステム監視状況をRI2N/USRのリンク故障・復活情報にフィードバックするシステムについて、概念設計を行った。

【研究項目４：クラスタ構築技術グループ】

① Soft Failure検知

本年度は昨年度に引き続き、故障発生器の開発を行った。発生可能な故障としては、現在までのところ、パケットを一定割合でのロスさせることができる。また、故障発生器によるシステムの性能へのオーバーヘッドを、アプリケーションベンチマークを用いて実験・評価し、オーバーヘッドは十分小さいことを確認した。さらに、クラスタ全体の故障を一元管理するためのリモート制御機構を実現した。

② 耐故障性MPIの実現

近年いくつかの耐故障性 MPI が実装され始めているが、それらは実行環境ごとに変わる Recovery Model への対応が困難である。我々が提案するCuckooMPIでは、Fault Model に対しRecovery Protocol をコンポーネント化することにより、実行環境に適した耐故障性 MPI を容易に構成である。また CuckooMPI は依存性の高い耐故障性機能も可換であるよう、コンポーネント分けされている。MPICH をベースとしてプロトタイプを作成し、実行時性能を測定したところ、その性能は MPICH の±2%程度であることを確認した。

③ チェックポインティングの高速化

昨年度実装したプロトタイプの投機チェックポインタを使用し、仮想的な並列環境で評価を行った。その結果アプリケーションによってはチェックポイント時間を最大41%程度削減した。その一方で多くのアプリケーションではほとんど効果が見られなかったものの、オーバーヘッドも非常に小さかった。これは投機ミスが少なく、アルゴリズムが一般的な状況で有効か、少なくとも実行時間を悪化させることがなく、有用な拡張であることを示唆していると考えられる。

④ 仮想機械を用いたマイグレーション可能なMPI

仮想機械の実装であるXenを用いて、MPIプロセスが動くゲストOSごと他の計算機にマイグレーションを行い、実際に計算が再開されることを確認した。また、Xen上のMPI計算でのオーバーヘッド、マイグレーションのコストを計測し、そのいずれもが十分軽微であることを確認した。また、Xenによるマイグレーションを用いてクラスタ上で負荷分散を行えるシステムのプロトタイプを実装し、実際に遊休ノードへの負荷分散が行われ、計算の効率化可能であることを確認した。

【研究項目５：プログラミング技術グループ】

① 実行モデルに含まれる異種コスト（計算、タスク間通信、タスク内通信）を統合して実行コストとする枠組みを構築し、通信負荷が大きいタスクの実行コスト予測精度を向上した。またプロファイル結果を実行モデルに反映してモデルを精緻化する機構の構築するとともに、MegaScript中のプロファイル指定の記述方式を拡張し、プロファイル対象の指定をより簡便なものとした。実行コスト予測法とプロファイル結果によるモデル精緻化の結果、従来は最大72%あったモデル誤差を23%にまで低減した。

② タスク並列実行のランタイム機構について、タスク間通信機能の拡張とタスク内並列実行への対応の機能追加・拡張を行い、記述容易性と実行性能の向上を図った。ランタイム機構の改良によりMPIを用いたタスク内並列実行が可能となり、またタスク間通信機構の改良により通信性能が５倍以上向上した。

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、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

(1) 論文発表

○ 堀田義彦，佐藤三久，朴泰祐，高橋大介，中島佳宏，高橋睦史，中村宏．プロセッサの消費電力測定と低消費電力プロセッサによるクラスタの検討．情処論ACS，Vol. 45，No. SIG11 (ACS7)，pp. 207-218，October 2004.

○ 藤田元信，田中慎一，近藤正章，中村宏．ソフトウェア制御オンチップメモリにおけるスタティック消費電力削減手法．情処論ACS，Vol. 45，No. SIG11 (ACS7)，pp. 219-228，October 2004.

○ Yoshihiko Hotta, Mitsuhisa Sato, Taisuke Boku, Daisuke Takahashi, and Chikafumi Takahashi. Measurement and Characterization of Power Consumption of Microprocessors for Power-Aware Cluster. In *COOL Chips VII*, April 2004.

○ Kenichi Kurata, Vincent Breton, and Hiroshi Nakamura, Secret Sequence Comparison in Distributed Computing Environments by Interval Sampling, In *CIBCB 2004*, pp.198-205, October, 2004.