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）

「低電力化とモデリング技術によるメガスケールコンピューティング」

1.研究実施概要

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において担当する要素技術に関して本格的な設計・検討を行い、主な成果として以下を得た。

(1) プロセッサグループ：消費電力削減効果の評価、低電力化アルゴリズム、ディレクティブベースコンパイラ
(2) コンパイラグループ：電力測定システムと低電力化効果の実測、プロトタイプMegaProtoの仕様策定
(3) ネットワークグループ：RI2Nの高バンド幅機能実装・評価、耐故障アルゴリズム
(4) クラスタ構築技術グループ：実行モデル構築およびコスト予測機構試作、耐故障基本機能試作
(5) プログラミング技術グループ：タスク並列スクリプト言語MegaScript仕様策定、静的実行モデル仕様策定

2.研究実施内容

【研究項目１：プロセッサグループ】

(1) メモリトラフィック削減、およびウェイ選択アクセスによる低電力効果の定量評価

提案するメモリアーキテクチャがデータアクセスの局所性を有効に活用することでオフチップメモリへのアクセスを削減することができ、その部分の消費電力を削減できること、および、オンチップメモリへのアクセスにおいてもキャッシュの全てのウェイをアクセスする必要が無くなることからその部分の消費電力を削減できる事を、

NAS Parallel benchmarkおよび、SpecFP95 benchmarkからいくつかのプログラムを用いて、定量的に評価した。その結果、メモリアクセスに要するエネルギーの40%-75%を削減できることがわかった。

(2) 低電力化方式のアルゴリズム化

(1)の低電力化のためには、オンチップメモリを有効に活用するような命令列を生成するコンパイルが必須になる。コンパイラグループと協調して研究を推進しているが、今年度は、高性能化のためのコンパイルアルゴリズムを完成した。(1)の評価結果は、このコンパイルアルゴリズムを用いた場合であり、さらに、低電力化のためのアルゴリズムを追加すれば、さらなる低電力化が図れる可能性がある。

(3) ディレクティブベースコンパイラ実装

(2)ではアルゴリズムを完成させたが、そのアルゴリズムは自動化されておらず、オンチップメモリのアーキテクチャに精通したプログラマによる人手変換が必要であり、従来のアーキテクチャとのプログラム上の互換性も失われる。この問題を解決するため、コンパイラグループと協調して、ディレクティブベースのコンパイラを開発した。これにより、従来のアーキテクチャとのプログラム上の互換性を維持したまま、人手変換の労力を格段に減少することができた。

(4) ディレクティブ自動挿入

(3)で開発したコンパイラでは、品質の良いコードを生成するためには、オンチップメモリのアーキテクチャに精通した人が適切なディレクティブを挿入する必要があった。この問題を解決するため、オンチップメモリのアーキテクチャに精通していなくても、処理対象プログラムのデータアクセスの特徴のみをヒント情報として与えることで、(3)のディレクティブを自動的に挿入できるコンパイラのフロントエンドの開発に着手した。これもコンパイラグループと協調しており、現在のところ、ヒント情報の仕様を確定し、ヒント情報からディレクティブに展開する部分の試作が終了している。

【研究項目２：コンパイラグループ】

(1) 電力計測システム

ホール素子による電力計測システムを構築し、これまで、PC等に用いられているPetinum IIおよびPentium4、および低消費電力向けのXscale,Transmeta Cursoeについて、電力計測し、さまざまなプログラムによる電力特性について明らかにした。また、メモリ最適化による電力削減効果について定量的に評価し、既存の低消費電力プロセッサにおけるコンパイラによる低消費電力化について検討を行った。

(2) アーキテクチャ駆動コンパイル

ハードウェアのパラメータを可変にするコンパイラ構成の変更機構については、前年度に引き続き、検討実装を行っている。

(3) メモリ階層利用最適化アルゴリズム

メモリ階層を最適に利用するコンパイラアルゴリズムについて、プロセッサグループの検討に基づき、アルゴリズムの実装に必要なツールキットモジュールの作成を行った。また、クラスタ構築技術グループとともにプロファイルによる最適化についてOmniコンパイラツールキットによる支援を行った。

(4) プロトタイプシステム

プロトタイプMegaProtoの仕様設計の計画を１年早め、ノードプロセッサ候補の絞込みと基本ノードの方式設計を行った。その結果、消費電力をプロセッサあたり10W以下に押さえつつ、標準キャビネット実装時に1Uあたり16プロセッサ以上の高密度実装と、プロセッサあたり1GFlops・2Gbps以上の演算・通信性能が得られることが明らかになった。またこの仕様に基づき、平成１５年度中にMegaProtoの第一バージョンを構築できる見通しを得た。

【研究項目３：ネットワークグループ】

(1) 高バンド幅機能実装

Fast Ethernetにおける、マルチポート・ネットワーク・インタフェース（MP-NIC）を用いたソフトウェア制御による高バンド幅クラスタ向けネットワークRI2Nの実装と性能評価を行った。年度始めに予定していたマルチストリームの並列転送は予備評価のみとし、シングルストリームの並列転送を対象としたユーザレベルライブラリを構築した。ライブラリは基本的に既存のTCP/IPと互換性を保っており、マルチスレッド及びselectシステムコールの2種類の実装を行い比較した結果、Fast Ethernetでは後者の実装が有利で、かつ比較的小さな単位での並列転送で、最大4ポート並列までのスケーラビリティが確保できることを確認した。

(2) Gigabit Ethernet評価

研究（1）をGigabit EthernetによるMP-NICに展開し、プロセッサ性能に対してネットワークバンド幅がかなり高いケースについてのRI2Nの性能評価を行った。（1）の結果より、当初からselect法のみを対象としたが、予想通りFast Ethernetに比べると転送単位が性能に与える影響が大きいため、適当なプロトコル変更を動的に行う必要があることが確認された。また、プロセッサ上でのネットワーク処理が相対的に重くなるため、3本以上のマルチリンクを用いることは得策でないことが確認できた。

(3) 冗長リンク方式検討

RI2Nにおける冗長リンク方式に関する検討を行った。Select法が今後の実装対象となることがほぼ決定したため、その上でのダイナミックな耐故障性実現のための検討を行った。RAID-5型（常に冗長パケットを送出）と、故障検出対応型の2種類のアルゴリズムを検討した結果、後者を採択することとした。現在の高バンド幅リンク実装プロトコル上で流れるハンドシェークパケットを利用したハートビート検出法で故障を検知、代替経路にスイッチするアルゴリズムを採用することとした。

【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

モデルによる実行コスト予測は、静的なプログラム解析による情報をコードの実行時の実際の振る舞いで精緻化することで実現する。従って、予測システムには実行時プロファイルが必要であり、軽量かつ正確でなければならない。我々は種々のプロファイル手法を検討した結果、インストラメンテーションサンプリングによる手法を選択した。実際に同手法を予備実装し、評価した結果、コスト予測のための情報を軽量かつ正確に得られることを確認した。

(2) Omni/SCASH における動的負荷分散

メガスケールシステムでは checkpoint からリスタートされる際、タスクが checkpoint 時と同等性能のノードに割り当てられるとは限らない上、そもそもノード間性能に heterogeneity が存在する可能性がある。本研究項目では、Omni/SCASH に対して、実行時プロファイルに基づくループ再分割機能を実装し動的負荷分散を実現した。また、データのローカリティが性能に大きな影響を与えるようなプログラムで特に重要となる page fault counting に基づくpage migration 機能に必要な、counting code および実験的な page migration コードを実装した。

(3) 耐故障性を備えたMPIの実装

逐次プロセスの checkpointer と通信路の耐故障性の確保によりMPI プロセスのcheckpoint/restartを行うシステムを構築した。プロトタイプとして、通信路の耐故障性には Rocks ライブラリを用い、Coordinated checkpointingを行うことによって耐故障性を実現した。また、実装したプロトタイプに対して Nas Parallel Benchmark および High Performance Linpack を用いて性能評価を行い、実行時オーバーヘッドが全体の性能に対して妥当な範囲に収まることを確認した。

(4) クラスタの復旧インストールの自動化

クラスタの復旧インストール機構の一部である、設定のパッケージ化手法について、設定パッケージ間の依存関係生成用ライブラリを実装し、設定パッケージが含むパッケージ、ファイル間に生じる依存関係を正しく自動生成できることを確認した。

【研究項目５：プログラミング技術グループ】

(1) タスク並列スクリプト並列言語の設計

オブジェクト指向スクリプト言語Rubyをベースとしてタスク並列スクリプト言語MegaScriptを設計し、その言語仕様第１版を策定した。仕様の主な点は、(a) 実行される並列タスクはシステム定義クラスTaskのサブクラスとして記述すること、(b) Taskクラスのメソッドinterface をオーバーライドしてタスクの実行環境を定義すること、および (c) Taskクラスのメソッドbehavior をオーバーライドして実行モデル構築のためのメタ記述を行うことである。またメタ記述のためのbehaviorメソッドでの制御構文の意味と、実行コスト定義のためのシステム関数の仕様も定めた。

(2) 実行モデル構築

MegaScriptによるメタ記述に基づく、静的な実行モデルの仕様第１版を策定した。モデルは実行コストを表現するコストオブジェクトでラベル付けされたノードと、分岐確率でラベル付けされた有向枝からなるDAGであり、コストオブジェクトを実行時情報でinstantiateすることで初期的なモデルが構築される。また動的な精緻化のためのモデル操作の枠組みも定めた。

(3) 並列タスク実行機構

並列タスクは実行モデルを用いたコスト予測に基づいてスケジューリングされ並列実行されるが、その基本的な枠組みとして、タスクの並列実行機構とタスク間通信機構の試作を行った。またこれらの機構を用いてMegaScriptで記述されたタスクを実行するために、基盤となるシステムクラスとメソッドを設計して試作した。性能評価の結果、タスク間通信の性能に問題があることが判明したため、設計の見直しを行っている。

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

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.研究成果の発表

(1) 論文（原著論文）

○ M. Kondo, M. Iwamoto, and H. Nakamura. Cache Line Impact on 3D PDE Solvers.

Proc. the 4th International Symp. High Performance Computing (ISHPC 2002), Lecture Notes in Computer Science 2327, pp.301-309, May 2002.

○ 近藤正章、大根田拓、田中慎一、中村宏. ソフトウェア可制御オンチップメモリを用いた低消費電力化の検討. 並列処理シンポジウム JSPP 2002, pp.285-288, May 2002.

○ 高宮安仁、松岡聡. ユーザー透過な耐故障製を実現するMPIへ向けて. 並列処理シンポジウム JSPP 2002, pp.217-224, May 2002.

○ 外崎由里子、大野和彦、中島浩. 並列スクリプト言語(Perl)+の実装と設計. 並列処理シンポジウムJSPP 2002, pp. 241-244, May 2002.

○ Masaaki Kondo, Shinichi Tanaka, Motonobu Fujita, Hiroshi Nakamura. Reducing Memory System Energy in Data Intensive Computations by Software-Controlled On-Chip Memory, Proc COLP02 (Workshop on Compilers and Operating Systems for Low Power) in conjunction with PACT02, Sep. 2002.

○ Yoshiaki Sakae, Satoshi Matsuoka, Mitsuhisa Sato, and Hiroshi Harada. Towards Dynamic Load Balancing Using Page Migration and Loop Re-partitioning on Omni/SCASH. Proc. 4th European Workshop on OpenMP (EWOMP 2002), Sep. 2002.

○ Taku Ohneda, Masaaki Kondo, Masashi Imai, Hiroshi Nakamura. Design and Evaluation of High Performance Microprocessor with Reconfigurable On-Chip Memory, Proc. IEEE Asia-Pacific Conference on Circuits and Systems 2002, pp.211-216, Dec. 2002.

○ 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. ハイパフォーマンスコンピュータシステムシンポジウムHPCS2003, pp.47-54, Jan, 2003

(2) 特許出願

なし