

「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。

本年度は最終的な成果統合へ向けて、以下のような主要な成果を得た。

- (1) プロトタイプ：MegaProto/E の完成、基本性能評価
- (2) プロセッサグループ：スタティック／ダイナミック電力統合最適化の設計・評価
- (3) コンパイラグループ：プロファイル駆動型電力最適化の実装・評価
- (4) ネットワークグループ：RI2N/UDP の実装・評価、階層化 RI2N の設計
- (5) クラスタ構築技術グループ：耐故障 MPI Cockoo のチェックポイント改良・評価
- (6) プログラミング技術グループ：コスト利用スケジューラの実装・評価

2. 研究実施内容

【プロトタイプ】

プロトタイプの最終版である MegaProto/E の開発を完了し、総計 16 ユニット (256 プロセッサ) のシステムを完成させ各拠点に配置した。プロセッサを従来の Crusoe から Efficeon に変更するなどの改良を行った結果、NAS Parallel Benchmark の性能が最大 2.4 倍に、また High Performance LINPAK の性能も 1.7 倍に、それぞれ向上した。

この他、最終的な統合システムの構築に向けて、電力測定環境構築と電力評価、階層化 RI2N 実装に向けての設計、MegaScript 処理系の実装などを行った。

【研究項目 1：プロセッサグループ】

今年度は、熱を考慮したスタティック消費エネルギーとダイナミック消費エネルギーを統合した消費電力削減手法の構築と、そのためのコンパイラ最適化をおこなった。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ため、実行時間の減少が少ないとスタティック消費エネルギーは増大する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。また、スタティック消費電力は温度に強く依存する。そこで、プロセッサチップ上の使用メモリ量から実行時間を見積もり、さらにスタティック消費エネルギーとダイナミック消費エネルギーを実行時の温度を用いて定式化することで、温度に依存して利用するメモリ量を最適化するコンパイラ最適化アルゴリズムを開発した。このアルゴリズムは、プログラム中に定期的に温度を計測するコードと計測結果に応じて、上記定式化の結果を用いて利用するメモリ量を変更するものである。

今年度はさらに、実行時の動的な情報をも加味して、プロセッサの電源電圧と周波数を制御することで、より効果的に消費電力を削減する手法の確立をめざし、プロセッサが用意するハードウェアカウンタの実行時の値、プロセッサの周波数とプロセッサの実効性能の間の関係を統計的に学習する実験を行った。その結果、プロセッサの実効性能に大きな影響をあたえるハードウェアカウンタの種類はかなり限られることがわかった。

【研究項目 2：コンパイラグループ】

- ① 実証システム MegaProto/E において、環境整備を行い、電力性能の評価を行った。設計どおりの電力性能が得られていることを確認した。
- ② DVS(Dynamic Voltage Scaling)機能を持つ消費電力プロセッサ(Crusoe と Turion)を用いた PC クラスタにおいて、プロファイル取得機能を用いて、いくつかの周波数での並列プログラムの実行プロファイルと電力プロファイルを取得し、各プロファイル区間に対してシステム全体の電力性能（エネルギー遅延積）を最適化する動作周波数を選択するアルゴリズムを提案した。実際のクラスタにて、実験環境とシステムを構築し、並列ベンチマークプログラムを用いて、検証を行い、その有効性を示した。
- ③ DAG(Directed Acyclic Graph)のタスクグラフで表現される並列プログラムに対して、同期待ちに生ずる余裕時間(Slack Time)を利用し、周波数を変更して電圧を下げて実行することによって、実行時間を変えずに電力を削減する周波数制御アルゴリズムを提案した。実際のクラスタにて、実験環境とシステムを構築し、単純な master-worker プログラムや tree のタスクグラフで表現されるコレスキ分解などの並列プログラムを用いて、検証を行い、その有効性を明らかにした。
- ④ MegaNode プロセッサアーキテクチャで提案したオンチップメモリと類似の機能を持つ低電力プロセッサ SH-4 を用いて、オンチップメモリを用いることによるシステム全体の削減効果について、いくつかのオンチップメモリを用いて最適化したプログラ

ムの電力性能の評価を行った。これにより、実際のプロセッサにおける MegaNode アーキテクチャの電力性能の見込みを得ることができた。

【研究項目 3：ネットワークグループ】

- ① マルチスレッド実装による完全な耐故障性及び高バンド幅化機能を持つ RI2N/USR の実装方式として、UDP/IP によるユーザレベルライブラリとして RI2N/UDP を実装し、小規模クラスタシステム上で性能評価した。評価の結果、2 本の Gigabit Ethernet リンクに対し同システムを適用することにより、無故障時には TCP/IP による単一リンクのほぼ 2 倍の性能が得られ、1 リンク故障時には単一リンク並みの性能が得られ、さらに全てのリンクが故障した後もそのいずれかが復旧し次第、通信が再開されるという、RI2N/USR が目指した機能が完全に実現されたことが確認された。
- ② 当初の年度目標の一つであった、RI2N/USR 上への MPI ライブラリの適用は、RI2N/UDP の実装の完全化を優先したため方式設計レベルに留まった。しかし、MPICH の p4 ドライバを RI2N/UDP によって差し替えることにより MPI への適応は容易に実現できることを確認した。
- ③ Tagged-VLAN 技術に基づく階層化 RI2N 実現手法として、数千ノードまでのスケーラビリティを持ち、従来の超並列計算機と同等な fat tree 構造を PC クラスタでの Gigabit Ethernet の集合で実現可能とする VFREC-Net システムの実装と性能評価を行った。VFREC-Net によって、上位階層のスイッチとそれに接続されるリンク数を増やすだけで、本システム用に開発された特殊 Gigabit Ethernet ドライバとスイッチ上の VLAN 設定の組み合わせだけで、従来の IP レベルの全ての通信がバンド幅ボトルネックなしにノード数拡張に耐えることが確認された。
- ④ VFREC-Net を最終的に MegaProto/E 上に実装するため、MegaProto/E 上のオンボードスイッチの VLAN 構成を変更し、上位スイッチとの組み合わせにより 8 ユニット (128 プロセッサ) までの高バンド幅接続が VFREC-Net によって実現されることを確認した。

【研究項目 4：クラスタ構築技術グループ】

- ① MPI に対する Portable Fault Tolerant Framework の実装
本研究項目では、MPI が実行される環境に適した耐故障性 MPI が使用可能となる、Portable Fault Tolerant Framework、CuckooMPI を開発している。本年度は、CuckooMPI を昨年度までに実装していた MPICH-ch_p4 ベースから MPICH-ch_p4mpd ベースに変更した。これにより、並列チェックポインティングのアルゴリズムを容易に変更できるようになった。この変更による NPB Class S 実行時のオーバーヘッドは最大で 10% ほどであった。さらに、フランス L.R.I. の MPICH-V Team の元に滞在し、並列チェックポインティングのひとつである、Pessimistic Message Log based checkpointing を実

装した。

② チェックポインティング最適化

今年度は、実アプリケーションにおける投機チェックポイントの有効性を検証するためのシミュレータを構築し、それを用いた性能評価を行った。同シミュレータは、チェックポイントが正確なメモリ更新のタイミングを完全に知っているものとし、これにより投機ミスがない場合の投機チェックポイントの性能を評価することができる。同シミュレータ上で NPB をアプリケーションプログラムとして実験を行ったところ、従来のインクリメンタルチェックポイントに対して最大 32%の性能向上を得られることがわかった。

③ マイグレーション可能な MPI 実行環境の拡張

今年度も継続してマイグレーション可能な MPI 実行環境の研究を行った。新たに VPN を用いる事で、異なるネットワークへのマイグレーションを可能とし、実際にテストベッドを構築してその評価を行った。仮想計算および VPN の使用によりネットワークの性能低下が起こったが、タスク並列アプリケーションなどではその有用性を実証した。

【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の改良を行い、通信負荷が大きいタスクの実行コスト予測に通信ライブラリの特長や実行環境のネットワーク構造を組み込んだ。その結果、従来の方法に比べてモデル誤差が 1/5～1/400 に低減するという高い効果を得た。
- ② タスク並列実行のランタイム機構について、Phonix を用いた広域ランタイムにユーザレベル機能拡張の枠組みを導入し、極めて小さなオーバーヘッドでアプリケーション固有の通信フィルタリング、例外処理、耐故障機能などを実現できるようにした。
- ③ タスクスケジューラに対してコスト利用スケジューリング機能を追加し、異種の並列タスクからなる階層型並列プログラムに適用した結果、負荷の偏差が 1～2%という極めて良好なスケジューリングが可能となった。またコスト利用の階層型のスケジューラを開発し 10 万タスクのシミュレーション評価を行った結果、タスク間の依存性が小さい場合には理想値に対して 1%未満の精度で負荷が均衡することが明らかになった。
- ④ タスク並列実行のランタイム機構を MegaProto/E に実装し、最終的な統合システムのベースを構築した。

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、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、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、教授）
②研究項目：安価かつスケーラブルな高速・高信頼ネットワーク技術

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、教授）
②研究項目：実行モデル構築とモデルを利用した大規模コモディティクラスタ管理技術

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、教授）
②研究項目：メガスケールのタスク並列プログラミングと実行技術

4. 主な研究成果の発表

(1) 論文（原著論文）発表

- 中島浩, 中村宏, 佐藤三久, 朴泰祐, 松岡聡, 高橋大介, 堀田義彦. 高性能計算のための低電力・高密度クラスタ MegaProto. 情報処理学会論文誌：コンピューティングシステム, Vol. 46, No. SIG12 (ACS11), pp. 46-61, August 2005.
- 湯山紘史, 津邑公暁, 中島浩. タスク並列言語 MegaScript 向け高精度実行モデルの構築. 情報処理学会論文誌：コンピューティングシステム, Vol. 46, No. SIG12 (ACS11), pp. 181-193, August 2005.
- 丸山真佐夫, 津邑公暁, 中島浩. データ再演法による並列プログラムデバッグ. 情報処理学会論文誌：コンピューティングシステム, Vol. 46, No. SIG12 (ACS11), pp. 214-224, August 2005.
- Hiroshi Nakashima, Hiroshi Nakamura, Mitsuhsa Sato, Taisuke Boku, Satoshi Matsuoka, Daisuke Takahashi, and Yoshihiko Hotta. MegaProto: 1 TFlops/10 kW Rack Is Feasible Even with Only Commodity Technology. In Proc. ACM/IEEE SC'05 Conf., November 2005.