論文の内容の要旨

論文題目: 細粒度通信に基づく並列計算機アーキテクチャに関する研究
氏　　名: 児玉祐悦

大規模シミュレーションや大規模サーバなどにおいて、計算機に対する高性能化の要求は今後もますます増大していくと考えられる。高性能化のためには、プロセッサ単体の性能を向上させるとともに、並列化を行うことが不可欠である。しかし、単体性能にプロセッサ数をかけて得られるピーク性能に比べて、実際に並列計算機で実行した場合の実効性能は低いものとなっている。一般的に、この並列化オーバヘッドはプロセッサ数が大きくなると増大する傾向にある。これまでは、この並列化オーバヘッドを軽減するために、通信のスループットを増大させるとともに、計算時間と通信時間のオーバラップによる通信時間の隠蔽を図ってきた。しかし、十分な通信スループットおよび通信時間の隠蔽を実現するためには、一回の通信量を大きくするようにプログラムを書き変える必要がある。そのため、そのような変更が可能な粗粒度並列アプリケーションしか十分な並列化の効果が得られなかった。より容易に、かつより広範囲に並列オーバヘッドを軽減し、並列効率を高めるためには、複数のスレッドを切り変えてプロセッサの実行効率を高めるマルチスレッド技術が有効である。このマルチスレッド実行の効果を高めるためには、スレッド切り替えおよび通信セットアップにかかる時間を軽減して、より多くの時間を本来のスレッド実行に割り当てられるようにすることが重要である。

マルチスレッド技術には、1)逐次的に実行されるスレッドを適切に切り替えながら実行するシンプルマルチスレッド方式、2)サイクル毎に異なるスレッドからの命令を実行する細粒度マルチスレッド方式、3)複数の演算ユニットを持ち、複数のスレッドからの命令を同時に実行する同時マルチスレッド方式等がある。シンプルマルチスレッド方式は、スレッド切り替えのタイミングがより細かな単位(数十から数百クロックサイクル)を想定していることが多く、スレッド切り替えやスレッドスケジューリング、スレッド間の通信や同期などにハードウェアサポートが必要である。細粒度マルチスレッド方式は古くは HEP というマシンで提案され、現在 TERA 社(現 CRAY 社)により MTA というスーパーコンピュータとして実用化されている。同時マルチスレッドは比較的新しい方式で、スーパースカラ方式では演算ユニットを増加させても命令レベル並列性を十分には引き出せないことに対する解決策として提案された。

このようなマルチスレッド計算機を考える場合に、スレッドの生成をプログラム実行中に可能とする動的スレッド生成を基本としスレッド数に制限を設けないようにするか、プログラム開始時に静的に一定個数のスレッドを生成する方式とするかが大きな設計方針となる。スレッドコンテキストをハードウェアで保持する細粒度マルチスレッド方式や同時マルチスレッド方式では、スレッド数がハードウェアリソースによって制限されるため、スレッド数が増大する可能性のある動的スレッド生成を採用することは難しい。しかし、本論文では粗粒度並列処理ではうまく並列化できないようなプログラムを並列化して高速化することを目指しており、効率的な動的スレッド生成を可能とするとともに、スレッド数へは特に制限を設けないことを目指している。このため、スレッドコンテキストをメモリ上に持つシンプルマルチスレッド方式を基本として設計指針の検討を行う。また、本論文では千台規模の並列計算機を想定している。この場合、各ノードは 1 チップ+メモリ程度に集約する必要がある。本研究を開始した当時の 1 チップのゲート規模からは、複数スレッドコンテキストをハードウェア的に保持することは困難であった。このようにシングルチップにネットワーク機構とプロセッサ機構を登載するという制約から、各機構をシンプルな構成に限る必要がある。このハードウェアを頻繁に利用する機能に限り、それ以外の機能は最適化したプログラムにより処理するという考えは、RISC プロセッサアーキテクチャに通じるものであり、本アーキテクチャ設計でも重要な指針となっている。我々はこのような設計指針に基づき、世界に先駆けてシングルチッププロセッサによるシンプルマルチスレッド方式の高並列計算機 EM-4 を先に開発した。本論文はこの EM-4 の評価に基づき、さらにスレッド実効性能を高めるとともに、共有メモリアクセスの実行性能を飛躍的に向上させたアーキテクチャについての提案とその評価についてまとめたものである。

本論文では、レジスタの内容から直接パケットを生成する細粒度通信機構を用いて通信オーバヘッドを削減するとともに、パケットの到着に基づきハードウェアで直接スレッドを起動することによりスレッド起動オーバヘッドを軽減したマルチスレッドアーキテクチャを提案する。本アーキテクチャでは、直接リモートメモリアクセス機構、局所同期機構などをハードウェアでサポートすることにより、メッセージ通信型プログラミングから共有メモリ型プログラミングまで柔軟なプログラミングが可能である。その際、ハードウェアとコンパイラの協調による処理の最適化に着目した実装を行っている。例えば、スレッドの切り替えタイミングを通信レイテンシが起きる時点としており、コンパイラがその切り替えタイミングを自動的に検出し、スレッド切り替え時に本当に必要なレジスタのみを待避/復帰するという最適化を行っている。これにより、スレッド切り替えのオーバ

ヘッドを削減している。また、基本的なスレッドライブラリなどはハードウェアのサポートにより数命令で実現できるため、インライン展開を適用することにより、関数呼び出しのオーバヘッドを削減している。

本提案の細粒度通信に基づくマルチスレッドアーキテクチャを実装した 80 プロセッサからなるプロトタイプ計算機 EM-X を構築した。この要素プロセッサとして、ネットワーク機構とプロセッサ機構を含めてシングルチップ化したプロセッサ EMC-Y を ASIC により開発した。本プロセッサを 5 個登載するプロセッサボードを 16 枚接続して 80 プロセッサシステムを構築している。この他、ホスト計算機との接続を行うインタフェースボード、ビデオ信号の入力や計算結果の出力を行うフレームバッファボードを開発した。これらのボードにはパケット処理用に EMC-Y が登載されており、プロセッサボード間を接続するケーブルの間に挿入する形で容易にシステムの拡張が可能である。また、それらとの通信は、プロセッサ間の通信と同様に細粒度パケットを用いた低レイテンシかつ高スループットな通信が可能である。本プロトタイプ計算機では独自のプロセッサを用いているため、コンパイラを始めとするソフトウェア環境も独自に開発を行った。ブロック転送、バリア同期／リダクション処理、ブロードキャスト、実行トレースなどのライブラリを専用に開発することにより、プログラムからハードウェア機構を有効に利用可能である。

提案したアーキテクチャの有効性を示すために、開発したプロトタイプ計算機を用いて各種ベンチマークによる性能評価を行った。最初に、並列プリミティブの評価を行い、リモートメモリアクセスの静的レイテンシが平均 1.3 マイクロ秒、2 点間スループットが 1K バイト程度の小さなブロックで 35M バイト/秒と、低レイテンシと高スループットを両立していることを確認した。実行時レイテンシの評価では、全プロセッサが動作しネットワーク上に 200M バイト/秒のデータが流れている状況でも、直接リモートメモリアクセスのレイテンシが 2 マイクロ秒程度であり、実行時レイテンシが極めて低く抑えられていることを確認した。バリア同期の評価では、局所同期を組み合わせたソフトウェアによる実装であるにもかかわらず、80 プロセッサのバリア同期が 13 マイクロ秒で行えることを確認した。また、カーネルベンチマークとして、行列乗算による評価では、小さい配列サイズから高い並列性能を達成できることを示すとともに、ナップサック問題や三角方程式の評価では、これまで並列化が難しいと考えられていた問題に関しても並列性能を引き出せることを示した。さらに、より大きなマクロベンチマークとして粒子シミュレーション MP3D と radix ソート用いて EM-X の全体性能について評価を行い、マルチスレッド処理によるレイテンシ隠蔽の有効性や、細粒度通信による

ネットワーク負荷の平均化の有効性などを示した。

従来、並列処理おいて性能向上を図るために通信粒度を大きくして通信スループットを向上させる手法が多かったが、本研究の成果によれば、シンプルであるが命令実行パイプラインと密接に融合した適切なハードウェアサポートにより、細粒度な通信のままでもそのレイテンシを削減/隠蔽することにより並列処理性能を向上させることが可能であることを示した。細粒度通信では、わざわざ通信を粗粒度にまとめることが不要であり、より細粒度な処理でも並列処理効果が期待されるため、並列処理の適用範囲を拡大することが可能となる。また、粗粒度通信では各プロセッサが一斉に通信を行うために通信の衝突による性能低下が引き起こされるが、細粒度通信では通信が平均的に散らばるためにネットワークの負荷が平均化され通信の衝突の影響が軽減されるという利点も見られた。